Teacher Quality and Student Inequality

Richard K. Mansfield

December 28, 2010

Abstract

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 an exogenous mobility assumption, allows separate identification of each teacher's quality from both school quality and school sensitivity to teacher quality. A test of exogenous mobility finds scant evidence that teachers systematically transfer to schools where they are relatively effective. 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 teacher and school inputs contribute 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.

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. We acknowledge the North Carolina Department of Public Instruction for collecting and providing this information.

1 Introduction

Recent research using matched student-teacher data has confirmed that teaching quality plays an important role in producing student test score improvement in elementary and middle schools (e.g. Rockoff (2004), Hanushek et. al. (2005), Kane et. al. (2007)). This finding has 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, research to this point 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.

Thus, this paper aims to answer three 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? (3) Would the performance of such disadvantaged students improve substantially if they were taught by better teachers?

Most of the existing literature focuses only on within-school variation in teacher quality.[1] Those papers that do consider differences in teacher quality across schools either estimate an upper bound on the variance in teacher quality that also reflects differences in other school-level inputs,[2] or they suffer from limited identifying variation (either small samples of schools, or small samples of transferring teachers connecting each school).[3] None provide a rigorous treatment of the conditions necessary for identification and precise estimation of school average teacher quality.

Furthermore, to this point nearly all of the attempts to examine the impact of teacher quality have employed elementary or middle school test scores.[4] However, there are a number of ad-

[1] e.g. Rivkin et al. (2005), Rockoff (2004).
[2] e.g. Hanushek et al. (2005).
[3] e.g. Aaronson et al. (2007).
[4] These include Hanushek, Rivkin, and Kain (2005), Boyd et al (2007), Goldhaber et. al. (2007), Kane et al.

vantages to using high school performance data to study the impact of teaching quality. First, we have very little sense of how much the quality of teaching matters at the high school level. 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 a given class), they can be more precisely estimated.

While one developing 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 effect that teachers' classroom assignments have on the variability in student performance over a high school career. 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.

Finally, most of the existing literature has ignored the possiblity that teacher quality may be complementary to other school-level inputs.[8] By allowing the impact of effective teaching to depend on the school, we can examine whether underprivileged students are in school environments that would enable them to profit from better teaching if it were provided.

We 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.

(2007), Kane and Staiger (2008), Jackson (2010) and Rothstein (2010). Aaronson et al (2007) is the exception, employing scores from 9th grade.

[5]e.g. Lankford et al. (2002), Steele et al. (2010)

[6]e.g. Hanushek et. al. (2005), Boyd et al.(2007), Jackson (2009)

[7]e.g. Clotfelter, Ladd, and Vigdor (2006), Rothstein (2010)

[8]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 et al. (2006)).

Such rich data permits 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. Separate identification of teacher quality from both school quality and school sensitivity to teacher quality stems from a large network of nearly 3,000 teacher transfers, coupled with a testable exogenous mobility assumption. This specification allows a teacher's true teaching talent to remain one-dimensional (enabling a meaningful discussion about teacher quality), but his/her effective ability to raise test scores to vary across schools. Allowing for such complementarity between school and teacher quality creates scope for efficiency gains from reallocating teachers. Given estimates of each teacher's quality, we then characterize the way quality teaching is currently being allocated. The next three paragraphs summarize our answers to the three questions posed above.

First, consistent with previous studies, we 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 .17 student test score standard deviations, enough to move an average student from the 50th test score percentile to the 57th percentile. However, 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 .061 student test score standard deviations, holding other school-level inputs fixed. This is only enough to move an average student from the 50th test score percentile to the 52nd percentile. Moreover, variation in teacher experience across schools, while substantial, contributes almost nothing to across-school test score gaps. Our 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, we 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, we find that such disadvantaged students are only slightly more likely to take classes with the relatively ineffective teachers at their schools. Overall, only about 4% 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.

Third, we 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 allocation 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 our student background index by .17 standard deviations.

Finally, in light of the provisions of the federal "Race to the Top" program, we also use our estimated model to evaluate the impact of a teacher accountability policy that denies tenure to those teachers whose estimated effects on student test scores place them in the bottom 5% of their cohort. We find that projections of the efficacy of such a regime depend on the model specification chosen; allowing for complementarity between teacher and school inputs leads to a far more pessimistic conclusion about the viability of such a policy. This is due to diminished confidence about our ability to correctly identify the least effective teachers.

The remainder of the paper is structured as follows. Section 2 presents the educational achievement production function along with the assumptions required to justify its form. Section 3 discusses identification of the parameters of the function. Section 4 describes the data. Section 5 presents the estimation strategy. Section 6 discusses the treatment of sampling error in parameter estimates. Section 7 presents the estimated distributions of teacher quality, school quality, and school sensitivity to teacher quality. Section 8 examines how quality teaching is allocated within schools and quantifies its impact on the distribution of average test scores over students' high school careers. Section 9 examines the contribution to achievement inequality of the existing allocation of teachers and schools to students. In Section 10 we use the estimated distributions to examine the impact of two counterfactual market interventions. First, we explore the impact of implementing the efficient allocation of teachers to schools on both the level and variance of student test scores. Then we examine the potential impact of implementing a test-score based tenure evaluation system for teachers. Section 11 examines the nature of teacher mobility, with an eye to its impact on the validity of the estimates. A surprising result emerges: teacher transfer patterns only reveal faint evidence of discernable job ladder among public high

schools. Furthermore, we test and fail to reject the assumption of exogenous mobility. Section 12 interprets the findings and concludes.

2 The Education Production Function

Following Todd and Wolpin (2003), we assume that the test score 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, we assume that current and past inputs are additively separable.[9] We also assume that current teacher and school inputs are additively separable from current student, family, and peer inputs, and from the measurement error.[10] We obtain:

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

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

$$h(\mathbf{R_{it}}, \mathbf{S_{it}}) = \delta_{s(i,t)} + \phi_{s(i,t)t} + \gamma_{s(i,t)}[\mu_{r(i,t)} + d(ex_{r(i,t)}) + \nu_{r(i,t)t}] \tag{2}$$

The first two components, $\delta_{s(i,t)}+\phi_{s(i,t)t}$, capture the collective impact of all school-level conditions that affect student learning independently of teacher quality. $\delta_{s(i,t)}$ represents the average impact of these inputs over the sample period. It reflects factors such as average principal quality, safety of the neighborhood, and the quality of school facilities. For the rest of the paper, we refer to $\delta_{s(i,t)}$ as the quality of school s. $\phi_{s(i,t)t}$ represents the transient component of school inputs, and captures fluctuations in principal quality, crime waves, renovations of school facilities, etc. $\delta_{s(i,t)}$ is a parameter to be estimated, while $\phi_{s(i,t)t}$ will be an error component.

[9] 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.

[10] 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, 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 (2007)). This 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.

The third component, $\gamma_{s(i,t)}$, is a school-specific scaling factor that captures the extent to which a school's students respond to the quality of their teachers. The intuition behind the school sensitivity parameter is that some schools may offer learning environments that either amplify or mute the impact of a teacher on her students' performance. For example, one can imagine a school that features a strict and effective disciplinary policy and classrooms with excellent acoustics, but gives teachers very little guidance as to how to craft a lesson plan. The impact of attending this school on a student's achievement will depend to a large extent on the quality of teachers the school employs. Such a school would have a particularly high value of γ. At the other extreme, a school may offer such crowded classrooms or outdated textbooks that even a very good teacher cannot raise test scores substantially; no one is learning regardless of who teaches them. Such a school would have a low value of γ (and perhaps a low value of δ). Alternatively, one might imagine a high achieving school with pre-designed lesson plans linked to instructional videos or computer applications, which again would make test scores less sensitive to the ability of the teacher who merely monitors the workstations. This school might simultaneously have a high value of δ and a low value of γ.

Note that many underlying school-level factors may contribute to both δ_s and γ_s. For example, an overly lenient discipline policy might decrease both δ_s and γ_s. However, since this analysis is focused on the impact on achievement of the overall school experiences received by different students, we 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. We treat each γ_s as a parameter to be estimated, and we will generally refer to it as the sensitivity of school s to teacher quality.

Since the median of γ will be normalized to one, the next three components, $\mu_{r(i,t)}+d(ex_{r(i,t)})+\nu_{r(i,t)t}$, can be interpreted together as the quality of teaching that student i receives in period t, as measured by the impact the teacher would have on the student's test score in a neutral learning environment.

First, in recognition of research indicating considerable persistent unobserved heterogeneity in teachers' performance, each teacher is assumed to have his/her own baseline ability to impact test scores, captured by $\mu_{r(i,t)}$. $\mu_{r(i,t)}$ is treated as a parameter to be estimated, and we will generally refer to $\mu_{r(i,t)}$ as teacher quality in the remainder of the paper.

Second, other research indicates that a teacher's effectiveness increases substantially with at least the first few years of experience (Clotfelter et. al. (2007), Rivkin et al. (2005)). Thus,

the function $d(ex_{r(i,t)t})$, assumed to be common across teachers, captures predictable growth in teacher effectiveness with experience, $ex_{r(i,t)t}$.

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

Together, the expression $\gamma_{s(i,t)}(\mu_{r(i,t)} + d(ex_{r(i,t)}) + \nu_{r(i,t)t})$ 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.

This specification for the contribution of school and teacher inputs to achievement offers a number of desirable features. First, the joint distribution of unobservable persistent teacher quality (μ) and two dimensions of unobservable persistent school effects (δ and γ) is left unrestricted.

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 across schools, as well as over time, as experience accumulates.

A third desirable feature is that school and teacher quality are allowed to act as either complementary or substitutable inputs, with the data informing us as to the extent of complementarity. Such input complementarity creates the possibility of efficiency gains from potential policies that incentivize a reallocation of teachers to schools.

The North Carolina data we use provides scores from tests based on ten distinct high school subjects, several of which may be taken in the same school year. Thus, the education production function must be altered to make it course-specific.

To do this, we allow the functions that map all past inputs and current student, peer, and family inputs into contributions to test score performance to be specific to the subject being tested, so that $f(*)$ becomes $f^c(*)$ and $g(*)$ becomes $g^c(*)$. School quality and teacher quality inputs are assumed to be common to all subjects. Most of the resources associated with a school that one expects to substantially impact test scores, such as principal quality, building facilities, and the safety of the surrounding neighborhoods, do not vary by course. In the case of teacher quality, high school teachers are only permitted to teach the subjects in which they are certified. Thus, in practice, we only need to assume that a teacher is equally effective at teaching,

say, Algebra 1 and Algebra 2. Finally, to minimize the impact of different choices of scales for exams taken in different subjects, we also standardize each test score relative to the appropriate course-year-specific state distribution, and re-interpret Y_{ict} accordingly. Thus, we have:

$$Y_{ict} = f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) + \\ \delta_s + \phi_{st} + \gamma_s[\mu_r + d(ex_r) + \nu_{rt}] + e_{ict} \tag{3}$$

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, we instead narrow our 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 qualities $\{\delta_s\}$, the set of school sensitivities to teacher quality $\{\gamma_s\}$, and the profile of teacher growth with experience $d(*)$. Given this focus, we 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 we denote $\mathbf{\tilde{Y}_i^{t-1}}$). Similarly, we use a vector of observable student, family, and classroom characteristics, denoted $\mathbf{X_{ict}}$, as a proxy for current student, family, and peer inputs. The impact of using these proxies is discussed in the next section.

Thus, the specification we estimate is:

$$Y_{ict} = \mathbf{\tilde{Y}_i^{t-1}}\alpha_\mathbf{c} + \mathbf{X_{ict}}\beta_\mathbf{c} + \delta_s + \gamma_s[d(ex_r) + \mu_r] + \epsilon_{ict} \tag{4}$$

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

$$\epsilon_{ict} = (f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) - \mathbf{\tilde{Y}_i^{t-1}}\alpha_\mathbf{c}) + (g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) - \mathbf{X_{ict}}\beta_\mathbf{c}) \\ + \phi_{st} + \gamma_s(\nu_{rt}) + e_{ict} \tag{5}$$

3 Identification

As emphasized by Todd and Wolpin (2003) and Meghir and Rivkin (2010), among others, endogenous choice of inputs by students, parents, and schools represents a formidable obstacle to identifying the parameters of an education production function. To fix ideas, we propose the following timing of input choices. (1) Given the history of school, student, and family inputs up through grade 8 combined with expectations about future student and family inputs during high school, parents choose high schools for their children. (2) School administrators assign

teachers to courses and tracks within courses. (3) Students submit desired course schedules, and are matched to teachers/classes. (4) Teachers and schools supply their inputs. (5) Parents and students choose their current inputs. (6) Standardized tests are taken.

Given this timing, the first potential issue is the parent's endogenous choice of school. Suppose that, conditional on observed prior test scores and observed current student, family, and classroom inputs, knowledge of a student's school does not provide further information about a student's unobserved prior inputs, nor his current unobserved individual and family inputs:

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

$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|1(s(i,t) = s'); \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] =$$
$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|\tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; s' \in \mathcal{S} \tag{6}$$

Then excluding prior student inputs will not bias estimates of persistent school quality (δ_s).

In practice, this condition seems unlikely to hold exactly.[11] Thus, to the extent that students with unobservably superior prior inputs systematically sort into particular schools, the values of δ associated with these schools will also reflect the impact of these inputs. Importantly, however, since γ_s and μ_r are identified purely from within school variation in test scores, violation of this condition will not bias estimates of these parameters.

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.[12] 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 the student attends:

[11] Assumption 1 might hold, for example, if past test scores were a sufficient statistic for the past inputs that are relevant for current test score performance and current inputs depended only on a student's post-graduation plans, which we observe in $\mathbf{X}_{ict}$. It might also hold if there was very little residential sorting, so that even if past test scores and observable inputs provide little guide to a student's unobservable current and prior inputs, knowing which high school a student attends provides no clues about his ability or the kind of family he has. Alternatively, if there was extreme residential sorting and a strict hierarchy of middle school quality and student ability by neighborhood, then knowing a student's past test scores and his parent's educational level might uniquely identify his neighborhood, and therefore his high school. In this context, Assumption 1 would also hold, since knowledge of the high school a student attends is redundant information.

[12] Rothstein (2010), in particular, has revealed the severity of this problem at lower levels of schooling.

Assumption 2: Conditional Mean Independence of Students' Unobserved Inputs and Teacher Identity

$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|1(r(i,t) = r'), 1(s(i,t) = s'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] =$$
$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|1(s(i,t) = s'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}]$$
$$\forall\ r' \in \mathcal{R}, s' \in \mathcal{S} \tag{7}$$

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. If students are essentially randomly assigned to teachers within track, then conditional on knowledge of a student's school, knowledge of a student's teacher should only provide information about the student's unobservable prior inputs if it reveals information about the students track, and knowledge of track is informative about these inputs. Thus, the validity of Assumption 2 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 we discuss the data in Section 4, it is worth noting here that we 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 may be a reasonable assumption at the high school level.

However, athird potential endogeneity problem, perhaps of greater concern in this context, is also ruled out by Assumption 2. Namely, even if students are conditionally randomly assigned to teachers, students (and parents) may respond to the quality of teaching they receive by adjusting their own current inputs. For example, a student saddled with an ineffective teacher may be more likely to study the textbook harder or pay for tutoring services.

If persistent teacher quality (μ_r) entered the production function linearly, such input compensation would cause estimates of teacher quality to be muted in magnitude, and the variance in teacher quality to be underestimated.[13] In contrast, in the non-linear model analyzed here,

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

to the extent that such input compensation is common to students from the same school, it will instead be reflected in a smaller estimate of the school's sensitivity to teacher quality, γ_s. Indeed, another advantage of this particular specification is that 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. To the extent that input compensation is beyond a school's control, one can simply reinterpret γ_s to be the sensitivity of the school's students to teacher quality given the kinds of students (and parents) that attend the school. Note, though, that the estimated variance in teacher quality will still reflect variation in the ex post impact of teachers after the amount of compensation that takes place at the median school, rather than the variation that would exist in the absence of any input compensation.

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. However, 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.

The fourth potential endogeneity problem stems from the possibility that teachers choose the effort they make or the content they teach. We do not explicitly model teacher effort. Instead, we 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}. The second 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 our data help allay these fears.[14]

A number of existing articles have already shown that assumptions like those made above are enough to identify and consistently estimate the distribution of teacher quality within schools.[15]

[14]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 we 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.

[15]e.g. Hanushek et al. (2005), Kane and Staiger (2008)

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 a student's choice to transfer to a different school is not correlated with changes in unobserved inputs. The second is that schools and student transfers form a connected graph (with schools as vertices and transferring students as edges).

Consider a special case of the model above in which $\gamma_s = 1 \ \forall \ s$. Then the model in (4) collapses to a version of Kramarz et al (2008) in which teachers take the place of students. Hence, in addition to Assumptions 1 and 2, identification requires that (1) a teacher's choice to transfer to a different school 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) and (2) that schools and teacher transfers form a connected graph, meaning that any two schools in the network can be linked by some chain of transferring teachers.[16]

The sufficient mobility condition is easily satisfied for a large network of schools in my data. Discussion of the connectedness of the schools in my data takes place in Section 4. As we 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 3: Exogenous Mobility

$$\begin{aligned} &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred to } s' \text{ at some } t' <= t] = 0 \\ &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred from } s' \text{ at some } t' > t] = 0 \end{aligned} \tag{8}$$

where ϵ_{ict} is the composite error term defined in (5). Given its importance, we will revisit the exogenous mobility assumption and the implications of its violation in Section 11. Essentially, while there are several mechanisms by which the exogenous mobility assumption could be violated, we devise tests for the two that are most plausible, and we fail to reject the exogeneity of mobility in either case.[17]

[16]Jackson (2010), in parallel work, uses a similar identification strategy. The focus of his paper, however, is on teacher-school match quality rather than the contribution of the way teachers are allocated to educational inequality.

[17]Note, though, that several obvious forms of systematic teacher mobility do not constitute violations of Assumption 3. These include systematic transfer of teachers to better schools, increased probability of transfer among inferior teachers, or even systematic transfer of effective teachers to schools more sensitive to teacher

The intuition behind the need for exogenous mobility is as follows. In the context where schools are equally sensitive to teacher quality, a teacher's persistent quality, μ_r, can be identified relative to the other teachers at their 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. With only one linking teacher, we would need to assume that his/her ability to increase test scores is the same across the two schools in order to infer relative school quality from his/her students' relative performance at the two schools. Once relative school qualities are known, we can shift the within-school distributions of average student residuals appropriately to place the performance of teachers from all schools in a neutral school environment. However, with many linking teachers, relative school quality is identified by differences in the *average* performance of transferring teachers across the two schools. Thus, we need only assume that transferring teachers do not *systematically* perform better at one of the schools, so that the difference in the average performance of students taught by transferring teachers still identifies relative school quality. This will be true if Assumption 3 holds.

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 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 to a common second school, and the difference between the two teachers' average test score residuals is larger at the second school than the first, this will contribute to a relatively larger estimated γ value at the second school.

quality. This is because the average performance among students taught by transferring teachers at each school under such mechanisms is predictable given knowledge of the teacher and school. Knowing that the teacher who teaches a given student transferred to the school does not provide any further information about any component of the composite error in these contexts.

[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.

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. However, because we can infer relative sensitivity independently using the two sources of variation we just described, we can still place all teachers in a neutral school environment. Now, though, in addition to shifting the distributions of teacher performance to adjust for differences in school quality, we must also rescale teacher performance at different schools by relative sensitivity to teacher quality before we can compare the quality of teachers at different schools. A formal proof of identification, given the assumptions and conditions above, is provided in Appendix 1.[19]

One should also 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 at a school must be zero by construction, since school averages of unobserved student-level inputs will be captured by the school-specific intercept, δ_s, and every student must have been taught by *some* teacher. Consequently, if transfer decisions are unrelated to the *bias* in transferring teachers' quality estimates,[20] 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.

The proof in Appendix 1 makes clear that $d(*)$, $\{\mu_r\}$, $\{\gamma_s\}$, and $\{\delta_s\}$ are only identified up to scale. Hence, we normalize the experience profile so that $d(0) = 0$. We normalize γ_s so that the median across schools is one; each school's value of γ_s captures the school's sensitivity as

[19]One may believe that the true production function involves different rates of teacher learning at different schools due to greater peer collaboration or higher quality feedback from principals and parents, so that $d(ex)$ should be $d_s(ex)$. In fact, I conjecture that $\{\gamma^s\}$ and $\{d_s(ex)\}$ can be separately identified if there exists a 2-edge connected graph of transfers between the $\mathcal{S}$ schools, and each experience level is represented at each school. While numerous examples have led me to believe this claim, I have not yet managed to prove it. However, I have proved a similar claim requiring a slightly richer network of transferrers, a requirement that is easily met in the data. It defines an algorithm that forms a network of schools by first finding two schools that have two teacher transfers between them, then inductively adding additional schools that are connected to the existing network of schools by at least two additional transferring teachers. The relative parameters of all schools (and, by extension, teachers) that can be connected via this algorithm are identified.

[20]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.

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_s is normalized so that it captures the expected increase in test scores from attending school s relative to a school with average δ, under the assumption that the global average of true teaching quality is 0. Appendix 2 discusses normalization and interpretation of parameter estimates in more detail, including how the normalization is implemented given a set of raw parameter estimates.

Given identification, one can shift focus to the precision with which parameters can be estimated. The proof in Appendix 1 makes clear that if there are merely enough transferrers to connect the schools, then these transferring teachers need to teach a large number of students at each school in order for the parameters of interest to be precisely estimated. This is because any differences in the average performance of a transferring teacher's students at different schools attributable to sampling error from test score measurement error or unobserved student inputs will be wrongly attributed to differences in school quality and school sensitivity to teacher quality.[21]

More generally, though, the performance of each transferring teacher at each school she teaches at provides information about the relative quality and sensitivity of the school (in addition to her own quality). Thus, our ability to distinguish school quality from average teacher quality actually depends on the total number of students taught by all transferring teachers at each school being compared.[22] When a moderate number of transferring teachers connect each school, school qualities, school average teacher qualities, and school sensitivities to teacher quality can still be precisely estimated even if each transferring teacher only teaches a moderate number of students at each school.[23] As we will see in Section 4, this is perhaps the more relevant scenario for our context.

4 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

[21]Furthermore, to the extent that the school-year and teacher-year idiosyncratic error components have considerable variance, each teacher needs to teach for a large number of years at each school as well.

[22]It will also depend on the total number of years taught by transferring teachers if ν_{rt} and ϕ_{st} are practically important.

[23]Of course, the precision of the teacher quality estimate ($\hat{\mu}$) of a particular teacher will still depend on the number of students and years he or she teachers.

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.[24]

The data contain a large number of observable current student[25] and peer[26] 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 seventh and eighth 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.

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. We limited the experience profile to six cells because constructing the design matrix for the school-teacher-experience cell combinations in the first stage of estimation (see Appendix 4) is computationally intensive, and increases rapidly in the number of cells. Furthermore, prior research suggests that most teacher learning occurs in the teacher's first few years (Clotfelter et al. (2007)).

Our method of estimation 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, we use a fuzzy matching

[24]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.

[25]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. We 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.

[26]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.

algorithm to match each teacher-class to a student-class. Since the grade level, race, and gender of each student in the student-class is observed, grade totals and race-gender cell totals can be constructed for the classes in the student-level data and compared to the corresponding grade totals and race-gender cell totals of the classes in the teacher-level data.[27] 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. Appendix 3 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. 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 teacher transfers, and teachers have often taught hundreds of students. To be included in the sample, we required teachers to have taught at least 20 students. When we impose the restriction that the network be 3-edge connected, so that any two schools can be connected using three distinct chains of transfers that do not share any links,[28] 329 schools remain. In fact, the majority of the 329 schools are far better connected: 217 of them are connected to each other by at least 10 transfers. Figure 1a shows the distribution across schools of the number of connecting transfers. Figure 1b shows the distribution of exams administered across teachers in the sample. Figure 1c shows the number of students taught by each transferring teacher at the school at which he/she taught fewer students. The latter figure illustrates that while some teachers have only taught one class at a second school, many others have taught at least 100 students at multiple schools.

After dropping students with missing test scores, teachers who taught at unconnected schools, and test scores from classes that were unmatchable, we are left with 4,016,343 test scores from 855,238 students and 18,498 teachers.

4.1 Choosing a Test Score Scale

The test scores are scaled scores that we re-standardized to have zero mean and unit variance within each subject-year combination. Meghir and Rivkin (2010) have noted that if monotonic

[27] 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.

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

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 particular concern in our context, since our method relies upon pooling tests from different subjects and years. One response is to note that 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 our 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.

Alternatively, we can assume that the specification we have chosen is modeling the relationship between the educational inputs a student received and the impact of the learning fostered by these inputs on another (standardized) outcome of arguably greater interest than test scores (e.g. money-metric utility or wages). Cunha et al. (2010), for example, explicitly anchor their test scores to wages, so that each coefficient can be interpreted as the wage value of the learning engendered by a given educational input. While we do not have an alternative outcome of interest to which our test scores can be anchored, we can interpret the production function we use as a misspecified version of the true function, with the parameters we estimate being approximations of the true parameters associated with the outcome of interest.

To make this argument precise, suppose that, for example, the money-metric utility value, denoted Y^M, of the learning fostered by our educational inputs after standardizing across students is appropriately modeled as $Y^M = w(*; \theta^M)$, where $w(*)$ is the specification in equation (4), and θ^M are the production function parameters of interest. Since a student's test score, denoted Y, is an alternative measure of student learning, it can be modeled as a monotonic transformation of $w(*)$, $u(w(*))$. Note that if $u(*)$ were an affine transformation, then standardizing test scores to be zero mean and unit variance would imply that *standardized* test scores bear the same relationship to inputs as *standardized* utility values, since standardizing eliminates the impact of any affine transformation. More generally, standardizing essentially amounts to using an affine approximation of the true transformation $u(*)$. Thus, to the extent that $u(*)$ can be well-approximated with an affine transformation, parameter estimates from estimating $Y = w(*; \theta)$ should closely approximate the true parameters associated with the outcome of interest, $\{\theta^m\}$.

While we have no way of verifying the extent to which $u(*)$ can be well-approximated by an affine transformation, we can detect evidence of floor and ceiling effects. These are particularly extreme forms of nonlinearity in test-score scaling where test scores reveal no information about the relative learning of a substantial part of the ability distribution. Because the scaling transformation was performed using total correct answers rather than particular response patterns, observed distributions of standardized scores would still reveal evidence of floor or ceiling effects in the tests, if they exist. Appendix 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. The appeal of this second interpretation of our results, however, clearly depends on one's preferred outcome measure, combined with one's belief about the relative value for that measure of the kind of learning facilitated by teachers at different parts of the achievement distribution in high school courses.

5 Estimation

From Section 2, the model to be estimated is:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}} \alpha_{\mathbf{c}} + \mathbf{X_{it}} \beta_{\mathbf{c}} + \delta_s + \gamma_s [d(ex_r) + \mu_r] + \epsilon_{ict} \tag{9}$$

We estimate equation (10) via non-linear least squares. Suppose that there are N student test scores associated with R teachers in S schools. In addition, there are K student- and classroom-level covariates and L prior test scores associated with each dependent variable test score, as well as J teacher experience cells. Then we need to minimize N squared deviations over K + L + 2S + R + J parameters. Since students' test scores partly reflect unobserved inputs and measurement error, a large number of student test scores are needed for each teacher in order to precisely estimate his/her quality. Also, a large number of schools is desirable in order to get a broad sense of how teaching quality is distributed across the state. Finally, a large number of transferring teachers is necessary to distinguish school quality from average teacher quality, and a large number of teachers at each school is necessary to get an accurate picture of the average teacher quality at each school. Thus, estimating the model is a daunting computational task: we employ over four million test scores and estimate nearly twenty thousand parameters. Fortunately, there are a few shortcuts that ease the computational burden considerably. Appendix 4 describes in detail the methods used to estimate the model.

Analytical asymptotic standard errors are calculated for all parameters. Standard errors are

clustered at the teacher-year and school-year levels, with the variance of the teacher-year and school-year components restricted to be common to all teacher-years and school-years, respectively. In addition, the variance of the test-score level error component (e_{ict}) is assumed to be homoskedastic. In order to make estimation of the variance-covariance matrix computationally feasible, the calculation needed to be broken down into several pieces. Appendix 5 describes the details of standard error computation.

6 Measurement Error

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 components that make up ϵ_{it}. To distill the true variance in teacher quality, we follow the approach used by Aaronson et al. (2007). First, we define each estimated teacher fixed effect $\hat{\mu}_r$ as the sum of the teacher's true quality and an uncorrelated error component: $\hat{\mu}_r = \mu_r + \xi_r$. Then the sample variance in estimated teacher quality can be decomposed as:[29]

$$\frac{1}{R}\sum_r(\hat{\mu}_r)^2 = \frac{1}{R}\sum_r(\mu_r)^2 + \frac{1}{R}\sum_r(\xi_r)^2 \tag{10}$$

Thus, one would like to estimate the variance in true teacher quality as:

$$\hat{Var}(\mu_r) = \frac{1}{R}\sum_r(\hat{\mu}_r)^2 - \frac{1}{R}\sum_r(\xi_r)^2. \tag{11}$$

ξ_r is not observed, but

$$\frac{1}{R}\sum_r(\xi_r)^2 \approx \frac{1}{R}\sum_r E[(\xi_r)^2] = \frac{1}{R}\sum_r(sd(\xi_r))^2, \tag{12}$$

so we estimate the error variance component using the standard error estimates for each teacher:

$$\hat{Var}(\mu_r) = \frac{1}{R}\sum_r(\hat{\mu}_r)^2 - \frac{1}{R}\sum_r(\hat{sd}(\xi_r))^2 \tag{13}$$

[29]We treat the set of teachers in North Carolina in my data to be the population of interest, and do not adjust for sampling error in the set of teachers observed. The variance in true teacher quality we calculate can thus be interpreted as the variance in the true quality of teachers teaching in my data, which covers all teachers in nearly all high schools in North Carolina over a ten year period.

We use the same technique to estimate the true variance in δ_s, γ_s, and school average teacher quality $\overline{\mu}_s$.[30]

By assuming that the errors and true qualities are distributed normally, we can then use these estimates of true variance in combination with the standard error estimates to calculate Empirical Bayesian posterior means for each estimated parameter using the standard Bayesian updating formula, given an assumed prior distribution with mean zero and variance $\hat{Var}(\mu_r)$. This Empirical Bayes estimator minimizes mean square error.[31] We use the Empirical Bayes estimates in our evaluation of a potential test-score based tenure evaluation system in Section 10.

7 Results

7.1 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 1 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 8 fewer years of experience than a school at the 95th percentile. For percent 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 2, 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 2 reveals that school quality and average teacher quality in fact have very limited scope to explain average test scores differences across schools. Rows 2 and 5 show that the lion's share (92%) of the total variance in test scores is due to differences in observable

[30] Given the variance matrix of $\{\hat{\mu}_r\}$, we use the delta method to calculate standard errors of $\overline{\hat{\mu}}_s$ for each school.

[31] For teacher quality, the formula is:

$$\begin{aligned}\mu_j^B &= \hat{\mu}_j(\hat{Var}(\mu_r)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2) + E[\mu_r](\hat{sd}(\mu_j)^2)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2)) \\ &= \hat{\mu}_j(\hat{Var}(\mu_r)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2)), \text{ since } E[\mu_r] \text{ is assumed to be 0.} \end{aligned} \quad (14)$$

Since the $\hat{\gamma}$ distribution is nearly log-normally distributed, we calculate Bayesian posterior means for $\log(\gamma)$, then exponentiate.

and unobservable student and classroom characteristics and test score measurement error, while Row 7 shows that observable differences in student and classroom characteristics explain three-quarters of the between-school variance. Row 3 indicates that unexplained variation between school-teacher-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, school sensitivity to teacher 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 $\delta_s + \gamma_s(\overline{\mu}_s + \overline{d(ex)}_s)$ 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 (moving from the 25th percentile school to the 75th percentile school would move the same student from the 47th to the 53rd percentile).

7.2 Teacher Experience

The estimated values of the teacher experience profile, $\hat{d(ex)}$, are as follows: first-year teachers are .034 student-level standard deviations worse than second-year teachers, .058 worse than third year teachers, .082 worse than teachers in their fourth through sixth years, .107 worse than teachers in their seventh through twelfth years, and .132 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 et al. (2007)).

These numbers, combined with the large differences in average teacher experience across schools displayed in Row 4 of Table 1, 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, we 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. Row 5 of Table 1 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 .027 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.

7.3 The True Variance of Teacher Quality, School Quality, and School Sensitivity to Teacher Quality

Figure 2 displays the histogram of estimated teacher effects $\{\hat{\mu}_r\}$. While the standard deviation of $\hat{\mu}_r$ is .307, correcting for sampling error leaves an estimate of the true standard deviation in teacher quality of .174 student-level standard deviations. The density of teacher quality (μ_r) is also plotted in Figure 2, under an assumption of normality. An average student who receives a teacher at the 75th percentile of teacher quality can expect to move from the 50th percentile to the 55th percentile, while one who receives a teacher at the 95th percentile can expect to move up to the 62nd percentile, assuming test scores are distributed normally.[32] This is substantial, and generally in line with most estimates found in the literature at the elementary and middle school level.[33]

To examine how average teacher quality is distributed across schools, Figure 3 plots the mean value of estimated teacher quality at each school, weighting each teacher by the number of students he/she taught at that school, which we denote $\hat{\overline{\mu}}_s$. Applying the analogous measurement error correction yields an estimate of the true between-school standard deviation of teacher quality of .061.[34] By imposing normality, we can also plot the density of $\overline{\mu}_s$ (the curve in Figure

[32]This assumption is borne out by plots of the data. See Appendix Figures 1-3

[33]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.

[34]We use the delta method to account for correlation in sampling error in $\hat{\mu}_r$ across teachers in the same school when calculating $sd(\overline{\mu}_s)$.

3). The estimates indicate 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 (54th) 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.

Figure 4 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. Considerable variation in γ_s exists: a school whose sensitivity to teacher quality is at the 5th percentile of the true distribution of school sensitivity is expected to be about .47 times as sensitive to teacher quality as the median school, while one whose sensitivity is at the 95th percentile is expected to be about 2.11 times as sensitive as the median school.

Finally, Figure 5 displays the histogram of school quality estimates, $\hat{\delta}_s$, alongside a plot of the underlying true distribution of school quality, δ_s, under a normality assumption. The estimated "true" standard deviation in δ_s is .112. If we assume that there are no unobserved student inputs that cluster at the school level, then moving from a school at the 50th percentile of the δ_s distribution to the 75th (95th) percentile would increase an average student's expected test score from the 50th percentile to the 53rd (57th) percentile, all else equal.

Table 3 displays the raw and true variances of the key parameters of the model for the baseline specification (first four columns) as well as a specification in which each school is equally sensitive to teacher quality: $\gamma_s = 1$. This linear specification represents the standard in the literature.[35] Given the large estimated variance in γ_s in the baseline model, the extent to which the standard deviations in μ_r (.174 vs. .172), $\overline{\mu}_s$ (.061 vs. .073), and δ_s (.112 vs. .090) in the restricted model mirror those in the baseline model is somewhat surprising.[36] Thus, any differences in the magnitude of teacher quality estimates between this and other analyses do not seem to be driven primarily by the non-linear specification employed here, but rather by the different samples of students and teachers (high schools versus primary/middle schools), and standardized test designs (course-specific versus general math and reading).

[35] See, for example, Aaronson et al. (2007) or Boyd et al. (2007).

[36] A likelihood ratio test overwhelmingly rejects the hypothesis that $\gamma_s = 1 \; \forall \; s$.

8 Student-Level Variance in Average Teacher Quality

While the results indicate that differences in 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.[37] 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, we 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. We denote this measure by $\hat{\overline{\mu}}_i$.[38] Using the delta method to calculate standard errors for each student's average teacher quality, σ_i^{μ}, we can estimate the variance in student-level teacher quality as:

$$Var(\overline{\mu}_i) = Var(\hat{\overline{\mu}}_i) - (1/I)\sum_i (\sigma_i^{\mu})^2. \tag{15}$$

Among students who took five tests, a one standard deviation increase in average teacher quality (relative to the overall school average) corresponds to an increase in average teacher quality of .062. 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 .078 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 47th (53rd) 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, we 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

[37] Note that such non-random assignment of students to teachers need not bias my 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.

[38] The results are similar if we condition on six or seven tests.

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.[39]

For each student we 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, we 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, we find that the across-student standard deviation in teacher quality under random assignment is .066 test score standard deviations.[40] Thus, within-school variation in the average teacher quality experienced across students does contribute to performance heterogeneity, but the variance in average teaching quality we observe is actually less than we would expect under random assignment.

9 The Impact of School and Teacher Inputs on Achievement Inequality

Given knowledge of the underlying distributions of school and teacher quality, in this section we examine the extent to which the existing allocation of students to teachers and schools is harming disadvantaged students. First, we look at whether disadvantaged students are more likely to receive their schools' relatively ineffective teachers. Second, we look at whether the schools that disproportionately serve underprivileged students are actually the worst schools.

[39]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, rescaling 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.

[40]We 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, we 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.

Finally, we attempt to provide an aggregate measure of the contribution of school and teacher inputs to achievement gaps between advantaged and disadvantaged students.

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

One way to interpret the results from Section 8 is that the existing mechanisms for allocating teachers to classes are not contributing to inequality in test score performance. However, it is still possible that students in lower tracks are systematically receiving lower quality teachers, and that the impact of such an imbalance on the student-level variance we 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, we 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, we first form an index of student background, $(X_{ict}\beta + \tilde{Y}_i\alpha)$, by weighting student characteristics by how well they predict high school test score performance.[41] Then, for each student among the bottom 10% of this index, we compare the average estimated quality of the teachers that actually taught the student with the average estimated quality among the set of teachers who taught the subjects the student took in the years they took them in their schools. We find that such students received teachers that were only .011 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 .022 better than the average teacher they could have expected under random assignment, given the teachers teaching in their subjects at the time. We find that black students received teachers who were .006 test score standard deviations less effective then they could have expected under random assignment. Furthermore, if instead we use as our 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, these small differences disappear entirely. Thus, we find very little 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 minuscule fraction of the performance gaps we observe.

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

9.2 What Kinds of Schools Disproportionately Serve Underprivileged Students?

The estimated parameter distributions displayed in Section 7 allow us to examine whether the schools disproportionately serving underprivileged students are actually the worst schools. To this end, Table 4 provides the average values of $\overline{\hat{\mu}}_s$, $\hat{\gamma}_s$, and $\hat{\delta}_s$ 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.

The last row presents results for our most comprehensive measure of average student background, the average value of the index $X_{ict}\beta + \tilde{Y}_i\alpha$. We find that high schools whose average indices across students place them in the bottom quartile of schools have teachers who are only .037 student level standard deviations below average, while those in the top quartile have teachers who are only .018 standard deviations above average. Columns 5 and 6, which replace $\overline{\mu}_s$ with $\hat{\delta}_s$, display essentially the same patterns; the schools in the top quartile of the average student index distribution increase test scores by .076 student-level standard deviations relative to schools in bottom quartile.

Columns 7 and 8 replace $\hat{\delta}_s$ with $\hat{\gamma}_s$. They reveal an interesting result: schools whose 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 25% to 40% higher than the overall median school in the sample. Most tellingly, schools in the bottom quartile of the $X_{ict}B + \tilde{Y}_i\alpha$ index have median sensitivity of 1.38, while those in the top quartile have median sensitivity of .70.

9.3 The Aggregate Contribution of School and Teacher Inputs to Inequality of Opportunity

While Table 4 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, we examine the typical allotments of these inputs received by particular subpopulations of students. In Table 5 we calculate the average values of the various school and teacher factors we have estimated among students in the bottom and top deciles and quartiles of our student background index, $(X_{ict}\beta + \tilde{Y}_i\alpha)$, as well as for both black students and white students.

We find that those among the bottom decile of the student background index attend schools that are .020 test score standard deviations below average and are 13% more sensitive than the median school. They also receive teachers that are on balance .022 test score standard deviations below average. Overall, the high school environment they experience (as measured by $\delta_s + \gamma_s(\overline{\mu}_i + \overline{d(ex)}_i)$) lowers their test scores by .047 standard deviations, relative to a typical environment in North Carolina. Students in the top decile go to schools that are 12% less sensitive than the median school but whose average quality is .017 standard deviations above average. They are assigned teachers that are .035 standard deviations above average. Overall, the high school environment they experience raises their test scores by .061 standard deviations. The difference in the school and teacher inputs the two groups receive can only account for .108 of the 2.5 standard deviation difference in their test scores (4%). Thus, the way teacher and school inputs are allocated in North Carolina is in fact exacerbating the existing disparity in performance between the least supported and best supported students; however, such unequal treatment is only making a marginal contribution to what was already a massive difference in high school achievement.

The comparison between races reveals a similar pattern. Black students attend slightly below average schools and receive slightly below average teachers in those schools, so that the overall high school environment they face lowers their test scores by .029 standard deviations relative to the average student. However, given that their test scores are .548 standard deviations below average, it is clear that while North Carolina high schools are not helping to fight racial achievement inequality, they are certainly not a major source of the problem. Differences in typical high school environments only account for 6% of the black-white test score gap.

10 Counterfactuals

10.1 Efficient Allocation of Teachers to Schools

Given the nature of complementarity between school and teacher quality in the model, a positive assortative match in which the schools most sensitive to teacher quality are paired with the most

effective teachers maximizes expected state average test scores. Thus, in this section we estimate the impact on both the level and distribution of student performance from implementing this allocation relative to both the status quo and a scenario in which average teacher quality is equalized across schools.

A number of caveats are in order. First, we hold fixed the sorting of students to schools we observe in the data. The endogenous response of parents to changes in school average teacher quality might erode some of the gains our model predicts. Second, to the extent that our school sensitivities are actually reflecting input compensation by parents and students rather than features of the schools themselves, changing the allocation of teachers may involve considerable costs borne by students and parents at schools estimated to be insensitive to teacher quality to minimize the impact of the decrease in average teacher quality they face. We do not model such costs. Third, as we discuss in Section 11, the impact of the match between a teacher and a school may not be fully captured by the quality/sensitivity complementarity in our model. If teachers and schools are currently sorting on the basis of an additional match component, then reallocating teachers may lower the average quality of matches along this unobservable dimension, offsetting the gains from putting good teachers in sensitive schools. Thus, rather than claiming to make an accurate prediction of the gains from this extensive reallocation, this exercise simply aims to determine whether the complementarities estimated in the model are large enough to have practical relevance.

Our methodology is as follows. We first estimate the joint distribution of school average teacher quality, school sensitivity to teacher quality, school quality, average teacher experience, and the average student background index.[42] Assuming multivariate normality, we can take 400 draws from this joint distribution to approximate the status quo. For each simulated school, we then simulate 10,000 test scores from 2,000 students taught by 50 teachers, using estimates of the variances of the other inputs of the education production function specified in (4).[43] Then, we construct the efficient allocation by reallocating the best 50 teachers to the most sensitive school, the next best 50 to the next most sensitive school, and so on.

We find that the efficient allocation increases the mean test score by .094 student-level stan-

[42]We estimate the covariances in a manner analogous to the estimates of true variances in Section 6. We calculate the covariance between raw estimates of parameters (say $\hat{\gamma}_s$ and $\hat{\delta}_s$), then subtract the average across schools of the covariance of the sampling errors associated with γ_s and δ_s ($\frac{1}{S}\sum_s cov(\epsilon_s^\gamma, \epsilon_s^\delta)$).

[43]These include the variance in within-school teacher quality, the variance in the within-student and between-student/within-school components of both the observable student background index and the unobservable idiosyncratic error, and the variance in year-specific deviations from long run school quality and teacher quality.

dard deviations, and reduces the standard deviation in test scores by 4.6%. Furthermore, the average test score among students in the bottom 10% of the student background index ($X_{it}B+\tilde{Y}_i\alpha$) is .175 test score standard deviations larger under the efficient allocation than under the status quo allocation. By contrast, the average test score among the top 10% of the student background index only increases by .029 standard deviations.

Note that the efficient allocation does substantially raise the variance in effective teacher quality ($\gamma_s\mu_r$) across students, but this effect on test score variance is outweighed by the fact that students enjoying increased effective teacher quality tend to have low values of other inputs. Interestingly, merely equalizing teacher quality by randomly assigning teachers to schools has almost no effect relative to the status quo: the average test score increases by only .012 standard deviations, the variance decreases by only half of a percent, and the average test score among the bottom 10% of the student background index only increases by .028 standard deviations. This is a testament to the fact that teaching quality is already surprisingly equitably distributed across schools.

Given the drastic nature of the efficient reallocation and the fairly small efficiency gain, 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.

10.2 Teacher Accountability Using Student Test Scores: A Tale of Two Standard Errors

The recent adoption of the federal Race to the Top program has given states a strong financial incentive to make the test scores of their students part of a public school teacher's formal evaluation. The education production function we have estimated allow us to examine the feasibility of such an evaluation policy. Suppose, for example, that North Carolina implemented a teacher accountability system that denied tenure to those in the bottom 5% of the posterior distribution of teacher quality after four years of teaching. Interestingly, the projected efficacy of such a policy depends on which specification is used to calculate parameter estimates, and more importantly, standard errors. Using Empirical Bayesian posterior means and variances of

teacher quality from the baseline specification[44], we find that 7.4% of those denied tenure under the above policy would in reality be above average teachers. If the students assigned to be taught by the denied teachers were instead randomly allocated to the remaining teachers, those students could expect an increase in teaching quality of .17 test score standard deviations. The overall average test score would increase by .007 standard deviations, after adjusting for slightly larger class sizes (a tiny adjustment).

A pertinent feature of our non-linear production function is that school sensitivity to teacher quality is imprecisely estimated even with considerable data, and the quality of teachers in insensitive schools (γ_s near 0) is very difficult to discern. The resulting uncertainty about which teachers are actually of the lowest quality is reflected in the policy's fairly small payoff and non-trivial risk of unfairly removing effective teachers. However, if we construct the Bayesian posterior distribution of teacher quality using estimates from the linear specification instead, then the (assumed) lack of uncertainty about school sensitivity implies that teacher quality can be estimated quite precisely. Thus, we find that only 0.1% of teachers fired would actually be above average teachers, and that the students who would have been taught by the removed teachers can expect a .35 standard deviation increase in teacher quality, resulting in an overall average test score increase of .018 standard deviations.

Notice in Table 3 that while the estimated variance in underlying teacher quality is nearly identical across the two specifications, the amount of sampling error present in the quality estimates of individual teachers $\hat{\mu}_r$ is much smaller under the linear specification (compare the columns entitled "error variance" across the two specifications). This comparison shows that the reliability of estimates of teacher quality, and by extension, the practicality of test-score based teacher accountability systems, depends not only on the quality of the tests administered and the number of students a teacher teaches, but also on one's belief about the appropriate specification of the education production function. Indeed, the specification used in this analysis, while general relative to much of the previous literature, may still fail to capture important input interactions which would affect the reliability of teacher quality estimates. Thus, the probability of correctly identifying an ineffective teacher may be considerably smaller than any model specification used so far would predict. Until we learn more about how different inputs interact to produce student learning, we must be extremely careful about how we interpret estimates of a single teacher's quality.

[44]Construction of Empirical Bayes estimates of teacher quality was described in Section 6.

11 Endogenous Mobility

Consistent estimation of the parameters requires that teachers' transfer decisions are unrelated to the composite error, ϵ_{ict}. This is a strong assumption with important implications for the validity of the estimates.

Specifically, recall that Assumption 3 requires:

$$\begin{aligned} &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred to } s' \text{ at some } t' <= t] = 0 \\ &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred from } s' \text{ at some } t' > t] = 0 \end{aligned} \tag{16}$$

Substituting the components in (5) for ϵ_{ict} in (16), we observe that a systematic relationship between a teacher's transfer decision and any of these components would violate Assumption 3. However, we restrict attention to two mechanisms that are particularly plausible, and develop tests for each.[45]

11.1 Do Teachers Try to Escape from Declining Schools?

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.

To test this hypothesis, we re-estimate the model with school-year additive effects, so that δ_s is replaced with δ_{st}.[46] Then, for each transferring teacher r, let $\tilde{t}(r)$ be the last year they teach at the school they transfer away from (denoted s). We can compute the average value of $\hat{\delta}_{st}$ for

[45] For example, one alternative is that measurement error in test scores or unobserved student inputs is related to teacher mobility, so that teachers are more or less likely to move when the test scores their students receive less accurately reflect the students' true talent, or when their students are underprivileged in a way that prior test scores and observed inputs would not reveal. We expect mobility driven by this mechanism, to the extent that it occurs, to resemble one in which teachers instead move to schools where they are better matched; both imply that a teacher should seem relatively more effective post-transfer than pre-transfer. Since we test for movement toward better match quality below, we do not develop a separate test for this possibility. A second alternative, related to ν_{rt}, is that teachers systematically transfer when their own quality is about to increase or decrease (relative to the standard experience profile and their own average quality over time). This might occur, for example, if teachers systematically transfer from urban to suburban schools when they are ready to start a family, and this coincides with them having less time to devote to lesson plan preparation, which decreases their effectiveness. However, given that most schools exhibit a mix of tranfers in and transfers out (see the subsection on mobility imbalance below), it seems unlikely that certain schools are systematically staffed with transferring teachers who happen to be having their relatively ineffective years there (over and above what can be predicted based on experience).

[46] Identification of this model requires a connected graph of teachers to link each school-year combination. But since the majority of teachers stay at a given school from one year to the next, connecting school-years within a school is trivial. And we already have verified the existence of a connected graph between schools.

the school he/she left for the years during/before their exit ($t <= \tilde{t}(r)$) and for the years after they left ($t > \tilde{t}(r)$). If teachers are transferring away from schools that are about to decline, then the mean difference among these two measures across transferring teachers should be positive:

$$\frac{1}{|\tilde{\mathcal{R}}|}\sum_{r\in\tilde{\mathcal{R}}}(\frac{1}{|\tilde{\mathcal{T}}^B|}\sum_{t<=\tilde{t}(r)}\hat{\delta}_{st} - \frac{1}{|\tilde{\mathcal{T}}^A|}\sum_{t>\tilde{t}(r)}\hat{\delta}_{st}) > 0. \tag{17}$$

Likewise, for each transferring teacher, we can compute the average value of $\hat{\delta}_{s't}$ for the school he/she joined (denoted s') for the years before his/her arrival and for the years after his/her arrival. If teachers are transferring toward schools that are about to improve, then the mean difference among these two measures should be negative:

$$\frac{1}{|\tilde{\mathcal{R}}|}\sum_{r\in\tilde{\mathcal{R}}}(\frac{1}{|\tilde{\mathcal{T}}^B|}\sum_{t<=\tilde{t}(r)}\hat{\delta}_{s't} - \frac{1}{|\tilde{\mathcal{T}}^A|}\sum_{t>\tilde{t}(r)}\hat{\delta}_{s't}) < 0. \tag{18}$$

We perform these tests, and find, strangely that while the schools that teachers join do indeed perform .009 student-level standard deviations better after the teachers arrive, the schools teachers leave also perform .012 better after the transferrers leave. The small magnitudes and conflicting interpretations of these test statistics suggest that teacher mobility is generally not driven by changes in school quality.

11.2 Do Teachers Move to Schools Where They Are Better Matched?

The education production function employed to this point has assumed that the match quality between schools and teachers is fully captured by the complementarity between teacher quality and school sensitivity to teacher quality. However, there is a growing literature inspired by Abowd et al.'s (1999) decomposition of wages that models employer-employee sorting based on match quality and its implications for the interpretation of firm and worker fixed effects (see Lise et al. (2008) and Lopes de Melo (2009) for examples). Some of the issues raised in this literature do not apply in the present context, since the outcome we are decomposing (essentially, average test score residuals for school-teacher combinations) is a direct measure of productivity, rather than an equilibrium object. Nonetheless, research by Jackson (2010) suggests that teacher-school match components are large enough to be economically important, and that teacher mobility might be related to match quality. Thus, in this section we entertain the possibility that ϵ_{ict} contains a match component, κ_{rs}, beyond that which is captured by the interaction between

teacher quality and school sensitivity to teacher quality:

$$\epsilon_{ict} = (f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) - \tilde{\mathbf{Y}}_\mathbf{i}^{\mathbf{t-1}}\alpha_\mathbf{c}) + (g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) - \mathbf{X_{ict}}\beta_\mathbf{c}) + \phi_{st} + \gamma_s(\nu_{rt} + \kappa_{rs}) + e_{ict} \quad (19)$$

κ_{rs} captures the possibility that teachers may be idiosyncratically more or less effective at teaching at particular schools. For example, such a match component might reflect the extent to which a teacher's teaching strengths coincide with how the principal wants lesson plans to be organized, or classrooms to be managed. The existence of the additional match component complicates interpretation of the estimated parameters, $\hat{\gamma}$, $\hat{\delta}$, and $\hat{\mu}$. In particular, a teacher's estimated quality, $\hat{\mu}$ will reflect not just her true quality, μ, but also her match component at her school (or, for transferring teachers, a weighted average of their match components at the schools at which they worked).[47] Our estimate of the variance in teaching quality within schools will now reflect both the variance in true teaching quality, μ, and the variance in the teacher-school match component, κ_{rs}. Note, though, that the composite teacher quality estimates and the composite within school variance estimate are actually the relevant factors for examining the contribution of the current allocation of teachers to student performance inequality, since it is the combination of μ_r and κ_{rs} that determines how effective teachers are in the schools at which they are actually teaching.

To the extent that schools and teachers can identify their potential match quality during job interviews, a model of sorting in the spirit of Lise et al. (2008) or Lopes de Melo (2009) might predict that the average match component among teachers at their initial school would be positive. However, since we only compare each school's quality relative to other schools, and each teacher's quality relative to other teachers, the average match quality among schools or teachers will have no impact on our estimates. To the extent that a particular school is relatively good at identifying teachers who will be good matches during hiring, all their teachers will perform relatively well there, so this will contribute to a larger school quality estimate, $\hat{\delta}$. This is appropriate, since such hiring skill would be a persistent school-specific characteristic. Similarly, if all above average teachers teach relatively well at a school, this suggests a high sensitivity to teacher quality at that school, and is precisely what we are trying to capture with $\hat{\gamma}_s$.

[47] To see this, re-examine equation (24) in the identification proof in Appendix 1. The difference in two teacher's average student residuals at the same school will now reflect the difference in the quality of their matches with the school, $\kappa_{12,1} - \kappa_{11,1}$ in addition to the difference in their true persistent qualities, even after the teachers have each taught for many years at the school.

However, the introduction of κ_{rs} creates another mechanism by which Assumption 3 might be violated: teachers might systematically transfer to schools at which they are relatively better at teaching. Such movement toward comparative advantage would imply that mobility is not merely potentially disruptive churning, but progress toward efficient allocation of teachers to schools.

11.2.1 The Extent of Mobility Imbalance

Developing a direct test of movement toward better match quality is more challenging. However, a crucial insight is that the impact of this form of endogenous mobility on parameter estimates depends critically on the extent to which mobility is "balanced". We refer to a school's pattern of teacher transfers as "balanced" if the number of teacher transfers away from the school equals the number of teacher transfers toward the school.

To see the importance of balanced mobility, consider the following simplified example. Suppose there are only two schools, A and B, with the same quality ($\delta_A = \delta_B$), sensitivity to teacher quality ($\gamma_A = \gamma_B$), and average teacher quality ($\overline{\mu}_A = \overline{\mu}_B$). Suppose further that a set of teachers transfer from A to B because they are better matched at B than at A ($\kappa_{rB} > \kappa_{rA}$). If these are the only teachers that transfer between A and B, then each transferring teacher has had their relatively ineffective years at A and relatively effective years at B, so that the average test residuals of their students will be higher at B than at A.

Under the assumptions of the model, the difference in the average test score residual among these transferring teachers at the two schools identifies the relative qualities of the schools. Thus, to best fit the model to the data, $\hat{\delta}_B > \hat{\delta}_A$, so we will overestimate the quality of school B relative to school A. Furthermore, since we have underestimated $\hat{\delta}_A$ and overestimated $\hat{\delta}_B$, the model fits the average scores of non-transferring teachers by overestimating the qualities of those at school A, and underestimating the qualities of those at school B ($\hat{\overline{\mu}}_B < \hat{\overline{\mu}}_A$).[48]

However, suppose there exists a second set of teachers of equal size that transfer from B to A because they are better matched at A than at B ($\kappa_{rB} < \kappa_{rA}$), and that the average magnitude of comparative advantage $|\kappa_{rB} - \kappa_{rA}|$ is the same across the two sets of transferrers. Then the average test score residuals at school A among the entire set of teachers who transferred between A and B will be the same as the average test score residuals at school B, so that the relative

[48] If transferring teachers tend to be above average teachers, then the model will also overestimate the sensitivity of school B, though this will be muted, since the relative performance of the good teachers compared to the bad ones among the transferrers at each of the two schools will also provide (correct) information about relative school sensitivities.

school qualities and mean teacher qualities of school A and school B will not be biased. Thus, if mobility is fully balanced, movement toward better match quality will not bias estimates of average school quality, average school sensitivity, and average teacher quality across schools. Appendix 6 offers a more formal treatment of this insight.

On the other hand, suppose there is a clear job ladder among schools, so that less desirable schools generally lose transferring teachers to more desirable schools and replace them with novice teachers, while more desirable schools tend to replace retiring teachers with transfers from less desirable schools. Then, directed mobility may lead us to underestimate the quality of less desirable schools relative to more desirable schools, and overestimate the average quality of their teachers.

Thus, in order to gauge the possible bias introduced by directed mobility, we examine the extent to which teacher mobility is balanced in our data. The simplest method is to calculate the fraction of each school's associated transferring teachers who transferred out (rather than in) and examine the distribution across schools. This approach has a couple of potential drawbacks. First, when new schools are created in a district, teachers may be involuntarily reallocated by the district to the new school. Consequently, any new school in our sample will tend to have joiners make up an overwhelming fraction of their transferrers, and other schools in the district will have leavers make up a disproportionate fraction of their transferrers. However, such involuntary transferring is unlikely to represent the kind of targeted mobility we are concerned about. Thus, when examining the distribution of the fraction of transferrers who are leavers, we eliminate in-transfers to new schools in their first year, and out-transfers to that school from any other school in that year. We do the opposite for school closings.

A second potential issue is that when we only observe a small sample of transfers from each school we should expect a sizeable number of schools to randomly have nearly all of their transfers in or out, even if no job ladder exists. While such small-sample imbalance could still bias parameter estimates, it will do so for a random selection of schools rather than for a particular type of school.

Thus, we also simulate two counterfactual densities of the fraction of transferrers who are leavers at each school: one in which no job ladder exists, and a second in which a fairly strong job ladder exists. In the first case, we fix the number of transferrers at the level observed in the data for each of the 329 schools in our sample, and assume that each of those transferrers was equally likely to be a leaver or a stayer. This would be the case in the absence of a job ladder, if

schools' teaching forces are remaining the same size over time. For each teacher, we take a draw, θ_r, from a Bernoulli distribution with $p = .5$, and assign this teacher to be a leaver if $\theta_r = 1$. We then calculate the fraction of each school's simulated transferrers who are leavers (denoted f_s), and sort the schools by this fraction f_s to get $\{f^1, ..., f^{329}\}$. We repeat this 100 times to get $f_b^1, ..., f_b^{329}$ for $b \in 1, ..., 100$, and average across simulated samples to get $\overline{f}^1, ..., \overline{f}^{329}$.

The method for constructing the density is the same in the case of a fairly strong job ladder, except that the draws are taken from a Bernoulli distribution with a school specific value of p, p_s. p_s is uniformly distributed on the interval $[.3, .7]$, so that some schools tend to be net senders (those with $p_s > .5$) and some tend to be net receivers ($p_s < .5$). The most desirable and undesirable schools will act as the sender 30 and 70 percent of the time, respectively.

Both counterfactual densities are plotted along with the true density of f_s in Figure 6. The first thing to notice is that even with small samples of transferrers at each school, mobility is fairly balanced in the data: 56 percent of schools send between 40% and 60% of their transfers, and 82% send between 30% and 70% of their transfers. This suggests that for a large set of schools, endogenous mobility may not introduce bias into estimates of their quality and average teacher quality, relative to others in the balanced set.

Second, the true density and the ladder-less counterfactual density are nearly on top of each other, while the counterfactual density associated with a moderately strong job ladder has considerably fatter tails. A Kolmogorov-Smirnoff test cannot reject the hypothesis that the true and ladder-less densities are identical, but overwhelmingly rejects the hypothesis that the true and laddered densities are identical. This suggests that the transfer patterns we observe in the data are consistent with the absence of a job ladder. While this method makes clear that much of the imbalance in mobility we observe need not reflect a systematic job ladder, it may overstate the amount of mobility imbalance that we would expect in the absence of a job ladder. This could occur if, for example, districts try to equalize experience across schools, so that within-district transfers are only granted if an offsetting trade is available. In this case, modeling each transfer as an *independent* Bernoulli draw will overpredict mobility imbalance.

Consequently, we turn to a second source of evidence for a job ladder, the observable characteristics of schools who are net senders or receivers of transfers. If mobility imbalance is pure small sample noise, then schools who are strong net senders should serve similar kinds of students as schools who are strong net receivers. Thus, we calculate the average student background index ($X_{ict}\beta + \tilde{Y}^i\alpha$) for schools in both the bottom decile and top decile of the distribution of

the fraction of transferrers leaving. The strongest net receivers (bottom decile) had students who were predicted to score .05 test score standard deviations above average based only on their observable characteristics, while the strongest net senders (top decile) had students who were predicted to score .22 test score standard deviations below average. This suggests that schools serving underprivileged students are somewhat more likely to lose teachers to other schools.

Overall, we see a modest amount of mobility imbalance, and while much of it is attributable to noise due to a relatively small number of transferrers at each school, mobility patterns do provide some evidence of a job ladder, in which students serving disadvantaged students tend to be on the bottom rungs. However, most transfers seem to be driven by factors other than the academic readiness of schools' students.

11.2.2 Testing for Endogenous Mobility

Recalling the two school example above, the evidence of systematic mobility imbalance, while fairly weak, implies that movement toward comparative advantage may lead us to underestimate the quality of schools serving underprivileged youth, and overestimate the quality of the teachers at these schools. This will occur if teachers decide to transfer only if the school serves better prepared students *and* they are better matched at such schools.

Fortunately, the two school example also suggests a possible test for mobility driven by match quality. Because the consistency of parameter estimates associated with a set of schools exhibiting balanced mobility does not require Assumption 3, if teachers are moving to better matches, a given transferring teacher should have his relatively ineffective years when teaching at the school he transferred away from and his relatively effective years when teaching at the school he transferred toward, compared to his overall average performance. Thus, for each teacher that transferred between two schools exhibiting balanced mobility, we calculate her average test score residual among students taught before transferring, and among students taught after transferring ($\bar{\hat{\epsilon}}_r^{Before}$ and $\bar{\hat{\epsilon}}_r^{After}$, respectively). The test statistic is the average difference between these residual means across all transferring teachers connecting schools featuring balanced mobility:

$$\frac{1}{|\tilde{\mathcal{R}}^B|} \sum_{r \in \tilde{\mathcal{R}}^B} (\bar{\hat{\epsilon}}_r^{After} - \bar{\hat{\epsilon}}_r^{Before}) \tag{20}$$

We can interpret this statistic as the average increase in teachers' abilities to increase student test scores following a transfer.

Under the null hypothesis of exogenous mobility, the average difference between these residuals across all transferring teachers should converge to 0 as the number of transferring teachers gets

large. If we restrict the sample of transfers to those occurring between the 56% of schools whose fraction of transfers leaving was between .4 and .6 in our data, the value of the test statistic is .009, with a standard error of .012. If we define the balanced schools more restrictively to be the 31% of schools whose fraction of transfers leaving was between .45 and .55, the test statistic is .016, with a standard error of .021. Notice that if these point estimates do capture the average increase in match quality associated with a transfer, then they would place an approximate upper bound on the extent of downward bias in the school quality estimates and on the corresponding upward bias in average teacher quality estimates for the strongest net senders as a result of this type of endogenous mobility. This is because even most net senders have at least a few offsetting arriving transfers, and we would expect these teachers to have their relatively effective years at these schools, thus counteracting part of the bias.

While we fail to reject the assumption of exogenous mobility, we do not have the power to rule out that some movement is driven by match quality or other components of the error term ϵ_{ict}.[49] However, violations of exogenous mobility do not seem to be introducing significant bias into estimates of the differences in quality or average teacher quality between schools.

12 Conclusion

In contrast to the horror stories recounted in the popular media in which the least privileged students attend disorganized schools with ineffective teachers, we find instead that quality teaching is fairly equitably distributed across high schools in North Carolina. Furthermore, differences in teacher experience account for a minuscule fraction of the test score gaps observed between schools. Instead, 90 percent of the between school variation is explained by student characteristics and prior test scores, suggesting that some combination of student ability, family inputs, and primary/middle school inputs account for most of the differences in performance across schools.

Disadvantaged students, as indicated by low values of the predictive index $X_{it}\beta + \tilde{Y}_i\alpha$, do tend to be exposed slightly inferior high school environments. 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 most of public school teachers' salaries are funded

[49]Note, though, that there is no monetary incentive for teachers to transfer toward better match quality, since teacher salaries only depend on education, experience, and district-specific premia.

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 (Rockoff et al. (2008), Clotfelter et al. (2007)). 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.

However, transfer patterns also provide only faint evidence of an underlying job ladder, suggesting instead the possibility 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).[50]

Another explanation may lie in the fact that teachers are hired by districts rather than schools, so to the extent that transfers 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, we 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 .17 student-level standard deviations at a school of median sensitivity, enough to move an average student from the 50th to the 57th percentile of the state test score distribution. Furthermore, there are 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 most sensitive to teacher quality seem to be those serving students whose characteristics are usually associated with low performance. Thus, policies that incentivize effective teachers to teach in struggling schools would be likely to increase average test scores statewide in addition to the

[50]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 Jackson (2009) and Goldhaber et al. (2007).

clear equity payoff. We estimate that the efficient allocation of teachers to schools would increase average test scores by .09 current test score standard deviations while simultaneously lowering test score variance by 5 percent.

The observed variation in teacher quality seems at first glance to suggest further opportunities for efficiency gains via policies that use test-score based teacher quality estimates to screen out bad teachers. Indeed, preliminary estimates based on a linear specification indicate that removing the teachers with the lowest estimated performance after four years might result in modest student gains with very few mistake layoffs of above average teachers. However, a closer look reveals that this conclusion depends on the choice of specification. More precisely, parameter standard error estimates are sensitive to the extent to which inputs are allowed to interact in the model, and the projected efficacy of such a policy intervention declines as our confidence in our ability to correctly identify ineffective teachers erodes.

Finally, given the sizable variation in teacher quality within schools, we 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 we find that which teachers a given student happens to receive has a modest but non-negligible impact on his overall performance in high school, the variation in average teaching quality experienced across students is fully explained by random assignment of students to teachers within a school. We find that disadvantaged students are only slightly less likely to receive the relatively effective teachers at their high schools.

A couple of caveats merit mention. First, the validity of the results depends upon the validity of the conditional random assignment and exogenous mobility assumptions. While the conditional random assignment assumption is more plausible in a high school context and is less critical for across-school comparisons, we do not test this assumption. The exogenous mobility assumption was tested and not rejected in my data, but the power of our tests was not sufficient to provide full verification. However, the extent of endogenous mobility seems to be quite limited, and the pattern of mobility we observe (specifically, nearly balanced mobility at most schools) suggests that such possible endogenous mobility may introduce only very limited bias into the parameter estimates.

Second, 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.

A natural path for future research would be to link these results for high school achievement to both prior school and teacher inputs as well as subsequent educational and labor market outcomes in a fully integrated dynamic model of education production. Cunha et al. (2010) estimate such a dynamic model for both cognitive and non-cognitive skill production, but only consider student and parent inputs. While existing research has looked at the impact of schools and teachers on elementary and middle school achievement, to this point little research has examined whether the impact of particular primary schools or teachers persists through high school and beyond.[51] Thus, I am currently examining whether attending particular elementary and middle schools and receiving particular elementary and middle school teachers can substantially increase high school performance, and whether the schools and teachers that increase high school performance are the same as those that increase elementary and middle school performance.

Similarly, the literature is still developing on the impact of school and teacher quality at the high school level on later educational outcomes. A landmark paper is Card and Krueger (1992), who examine the impact on earnings of attending a high school with more desireable observable characteristics. In current work, Altonji and Mansfield (2010) exploit the stratified structure of three national longitudinal surveys to estimate approximate bounds on the impact of high school quality on high school graduation, enrollment at a four-year college, and adult wages. While they also find that differences in school quality represent a small fraction of the total variance in both high school achievement and later outcomes, these differences nonetheless translate into substantial impacts on high school graduation and college enrollment, since large numbers of students seem to be near the decision margin. Fortunately, recent improvements in administrative data collection and dissemination should soon provide an opportunity to explore the long-run impact of teacher quality as well as high school quality.

[51] The recent evidence from Chetty et al. (2010) using the Tennessee Star experiment is an exception. They find that the large contemporaneous impact of randomly assigned classroom quality within an elementary school (a composite of teacher and peer quality) fades considerably when measured by 8th grade test scores, but re-emerges for both college enrollment and earnings at age 27.

13 References

Aaronson, D., Barrow, L, Sander, W (2007). "Teachers and Student Achievement in Chicago Public Schools." *Journal of Labor Economics*. vol. 25 no. 1

Abowd, John, Francis Kramarz, and David Margolis (1999). "High Wage Workers and High Wage Firms." *Econometrica*. Vol. 67, No. 2, pp. 251-333.

Abowd, John, Robert Creecy, and Francis Kramarz (2002). "Computing Person and Firm Effects Using Linked Longitudinal Employer-Employee Data". Working Paper. March 2002. http://courses.cit.cornell.edu/jma7/abowd-creecy-kramarz-computation.pdf

Altonji, Joseph, and Richard Mansfield (2010). "The Contribution of Family, School, and Community Characteristics to Inequality in Education and Labor Market Outcomes." Unpublished Manuscript.

Boyd, Donald, Pam Grossman, Hampton Lankford, Susanna Loeb, and James Wyckoff (2007). "Who Leaves? Teacher attrition and student achievement." Working Paper 14022. Cambridge, MA: National Bureau of Economic Research.

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*. Vol. 24, No. 1, pp. 113-132.

Card, David, and Alan Krueger (1992). "Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States." *The Journal of Political Economy*. Vol. 100, No. 1, pp. 1-40.

Chetty, Raj, John Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan (2010). "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." Working Paper 16381. Campbridge, MA: National Bureau of Economic Research.

Clotfelter, Charles T., Helen F. Ladd, and Jacob T. Vigdor (2006). "Teacher-Student Matching and the Assessment of Teacher Effectiveness." Working Paper 11936. Cambridge, MA: National Bureau of Economic Research.

Clotfelter, Charles T., Helen F. Ladd, and Jacob T. Vigdor (2007). "Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects." *Economics of Education Review*. Vol. 26, No. 6. pp. 673-682.

Cunha, Flavio, James Heckman, and Susanne Schennach (2010). "Estimating the Technology of Cognitive and Non-Cognitive Skill Formation." *Econometrica*. Vol. 78, No. 3, pp. 883-931.

Goldhaber, Dan, 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'Brian, and Steven Rivkin (2005). "The Market for Teacher Quality" Working Paper 11154. Cambridge, MA: National Bureau of Economic Research.

Jackson, C. Kirabo (2009). "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation". *Journal of Labor Economics*. Vol. 27, No. 2.

Jackson, C. Kirabo (2010). "Match Quality, Worker Productivity, and Worker Mobility." Working Paper 15990. Cambridge, MA: National Bureau of Economic Research.

Kane, Thomas, Jonah Rockoff and Douglas Staiger (2007) "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City" *Economics of Education Review*. May 2007.

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

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*. Vol. 24, No. 1, pp. 37-62.

Lise, Jeremy, Costas Meghir, and Jean-Mark Robin (2008). "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*. Fall 2009, Vol. 4, No. 4, pp. 439-467.

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

Meghir, Costas, and Steven Rivkin (2010). "Econometric Methods for Research in Education." Working Paper 16003. Cambridge, MA: National Bureau of Economic Research. Prepared

for the Handbook of Education.

Rivkin, Steven, Eric Hanushek, John Kain (2005). "Teachers, Schools, and Academic Achievement." *Econometrica*. Vol. 73, No. 2, pp. 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. Vol. 94, No. 2. pp.2 47-252.

Rockoff, Jonah, Bryan Jacob, Thomas Kane, and Douglas Staiger (2008). "Can You Recognize An Effective Teacher When You Recruit One." Working Paper 14485. Cambridge, MA: National Bureau of Economic Research.

Rothstein, Jesse (2010). "Teacher Quality in Education Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*. February 2010. Vol. 125, No. 1. pp 175-214.

Steele, Jennifer, Richard Murnane, and John Willett (2010). "Do Financial Incentives Help Low-Performing Schools Attract and Keep Academically Talented Teachers? Evidence from California." *Journal of Policy Analysis and Management*. Vol. 29, No. 3, pp. 451-478.

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

14 Tables and Figures

Table 1: The Distribution of Teacher Credentials Across Schools

Selected Quantiles of the Distribution of School Means of Selected Teacher Credentials					
Credential	5%	25%	50%	75%	95%
Percentage of Teachers w/Masters or Other Advanced Degree	.110	.196	.259	.332	.455
Percentage of Teachers w/ National Board Certification	0	.013	.039	.067	.138
Percentage of Teachers Who Are Uncertified	.053	.102	.138	.187	.284
Years of Teaching Experience	8.67	11.13	12.65	14.23	16.75
Average Effective Teaching Experience: d(exp)	-.015	-.006	0	.006	.013

Table 2: 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 $Var(X_{ict}\beta_c + \tilde{Y}_i^{t-1})$	.514	.717	.569
(3)	Effective School and Teacher Quality $Var(\lambda_{srj})^*$	.047	.216	.052
(4)	Cov(Stu. Background, Eff. Sch./Tch. Qual.) $2 * Cov(X_{ict}\beta_c + \tilde{Y}_i^{t-1}, \lambda_{srj})$	.029	–	.032
(5)	Idiosyncratic Test Score Error $Var(\epsilon_{ict})$	.317	.563	.351
(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	.242	.065
(8)	Total School Quality $Var(\overline{\lambda}_s)^{**}$	.009	.097	.010
(9)	Cov(Avg. Stu. Background, Total Sch. Qual.) $2 * Cov(\overline{X_s\beta_c} + \overline{\tilde{Y}}_s^{t-1}, \overline{\lambda}_s)$	.010	–	.011

$^*\lambda_{srj}$ is the mean unpredicted test score of students taught by teacher r in school s while the teacher was in experience cell j (See Appendix 4). $\lambda_{srj} = \lambda_h = \delta_s + \gamma_s(\mu_r + d(ex_{rt})) + \omega_{srj}$. Thus, $Var(\lambda_{srj})$ consists of the combined contributions of school quality, school sensitivity to teacher quality, teacher quality, teacher experience, and the component of the idiosyncratic error that is between school-teacher experience cells.

Note that "Student Background" includes the impact of classroom peers, since average observable characteristics of classmates are elements of X_{ict}.

Table 3: Raw and Error-Adjusted Variances in μ_r, $\overline{\mu}_s$, γ_s, and δ_s: Baseline and Linear Specifications

	Baseline Model $\delta_s + \gamma_s(\mu_r + d(exp_r))$				Uniform Sensitivity $\delta_s + \mu_r + d(exp_r)$			
Parameter	Raw Var.	Error Var.	True Var.	True Std.	Raw Var.	Error Var.	True Var.	True Std.
Teacher Quality (μ_r)	.077	.047	.030	.174	.040	.010	.030	.172
	(.005)	(.004)	(.002)	(.004)	(.001)	(.000)	(.001)	(.003)
School Average Teacher Quality ($\overline{\mu}_s$)	.012	.008	.004	.061	.008	.002	.005	.073
	(.001)	(.001)	(.001)	(.009)	(.001)	(.000)	(.001)	(.005)
School Quality (δ_s)	.020	.007	.013	.112	.011	.003	.008	.090
	(.002)	(.000)	(.002)	(.009)	(.001)	(.000)	(.001)	(.006)
School Sensitivity to Teacher Quality (γ_s)	.787	.467	.320	.566	–	–	–	–
	(.086)	(.037)	(.057)	(.051)	–	–	–	–

Approximate standard errors are in parentheses. They were obtained using bootstrap samples from the combinations of $\{\hat{\mu}, sd(\hat{\mu})\}$, $\{\hat{\gamma}, sd(\hat{\gamma})\}$, or $\{\hat{\delta}, sd(\hat{\delta})\}$ 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 4: Average Schooling Inputs among Schools in the Top Quartile Versus Bottom Quartile of Various Student Characteristics

	Mean Student Characteristic		Mean Teacher Quality ($\hat{\overline{\mu}}_s$)		Mean School Quality ($\hat{\delta}_s$)		Mean Sens. to Tch. Qual. ($\hat{\gamma}_s$)	
	Bottom	Top	Bottom	Top	Bottom	Top	Bottom	Top
Mean 8th Grade Math Score	-.155	.548	-.032	.025	-.028	.021	1.33	0.70
Percent Black	.058	.608	-.002	-.032	.026	-.019	.735	1.32
Percent Hispanic	.010	.086	-.002	.020	-.005	-.018	1.05	1.00
Percent Eligible for Free Lunch	.139	.516	.004	-.029	.037	-.018	0.87	1.34
Stu. Backgr. Index ($X_i\hat{\beta} + \tilde{Y}_i^{t-1}\hat{\alpha}$)	-.382	.260	-.037	.018	-.042	.034	1.38	0.70

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 ($\hat{\overline{\mu}}_s$) 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 School Quality is the average value of estimated school quality ($\hat{\delta}_s$) 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.

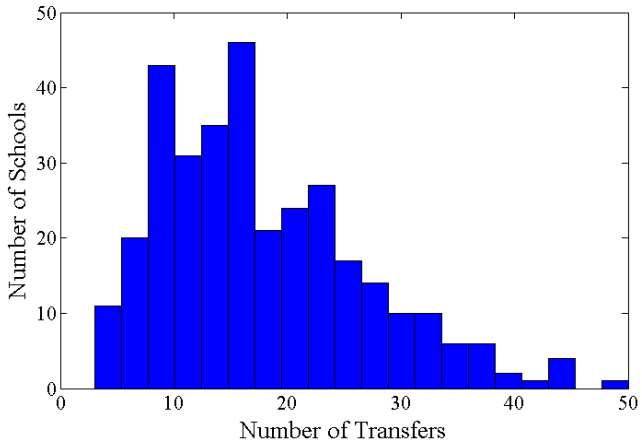
Stu. Backgr. Index is an index of student background composed of the predicted test score based solely on the student's current observable characteristics and test scores collected prior to high school.

Table 5: Average Schooling Inputs and Outcomes among Selected Subpopulation of Students, in Test Score Standard Deviations

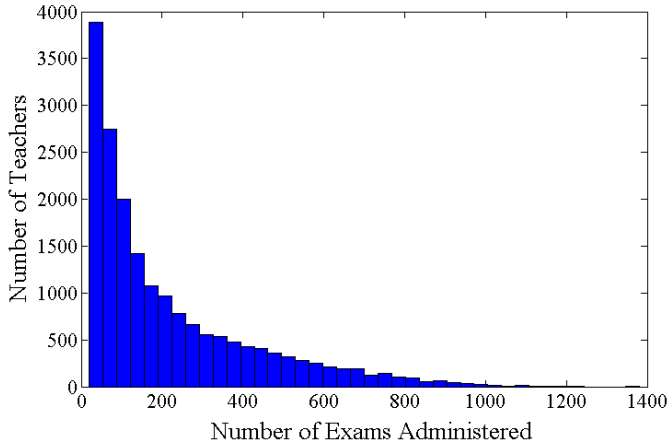
	Average Value of Input or Outcome Among Subpopulation					
Student Subpopulation	Test Score $(\overline{Y}_i)$	Total Tch. Qual. $(\overline{\hat{\mu}}_i)$	Within Sch. Tch. Qual. $(\overline{\hat{\mu}}_i - \overline{\hat{\mu}}_s)$	Sch. Qual. $(\hat{\delta}_s)$	Sens. to Tch. Qual. $(\hat{\gamma}_s)$	Total School Contribution $(\hat{\delta}_s + \hat{\gamma}_s(\overline{\hat{\mu}}_i + \overline{\hat{d}(ex)}_i))$
Student Background Index $(X_{it}B + \tilde{Y}_i^{t-1}\alpha)$						
Bottom 10%	-1.164	-.022	-.012	-.020	1.13	-.047
Bottom 25%	-.894	-.020	-.010	-.015	1.11	-.040
Top 25%	.974	.026	.015	.013	.90	.046
Top 10%	1.34	.035	.022	.017	.88	.061
Race						
White	.185	.007	.002	.007	.93	.017
Black	-.548	-.015	-.006	-.012	1.11	-.029

Figure 1: A Graphical Depiction of the Strength 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

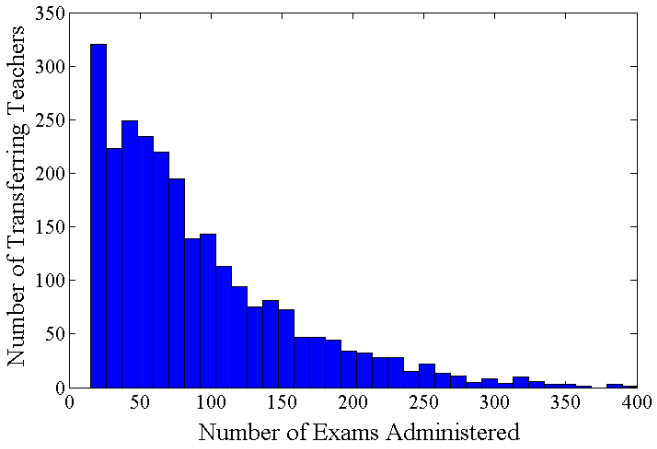


Figure 2: Individual Teacher Quality Estimates ($\hat{\mu}_r$) and the Underlying Density of Teacher Quality (μ_r), after Correcting for Sampling Error

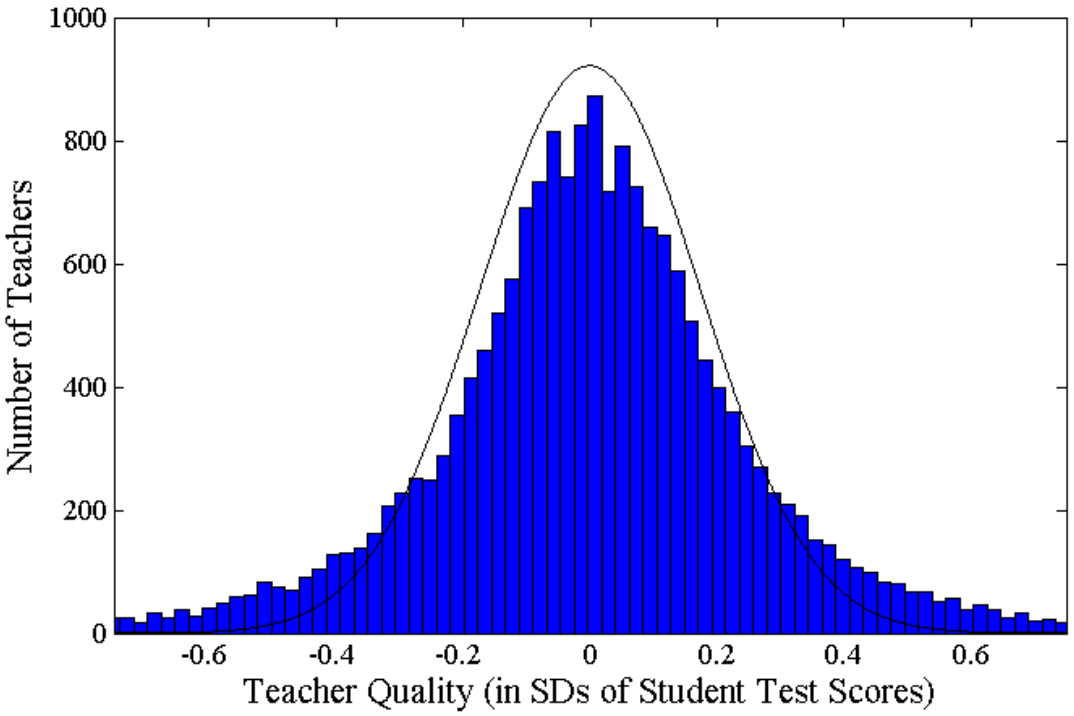


Figure 3: School Average Teacher Quality Estimates ($\hat{\overline{\mu}}_s$) and the Underlying Density of School Quality ($\overline{\mu}_s$), after Correcting for Sampling Error

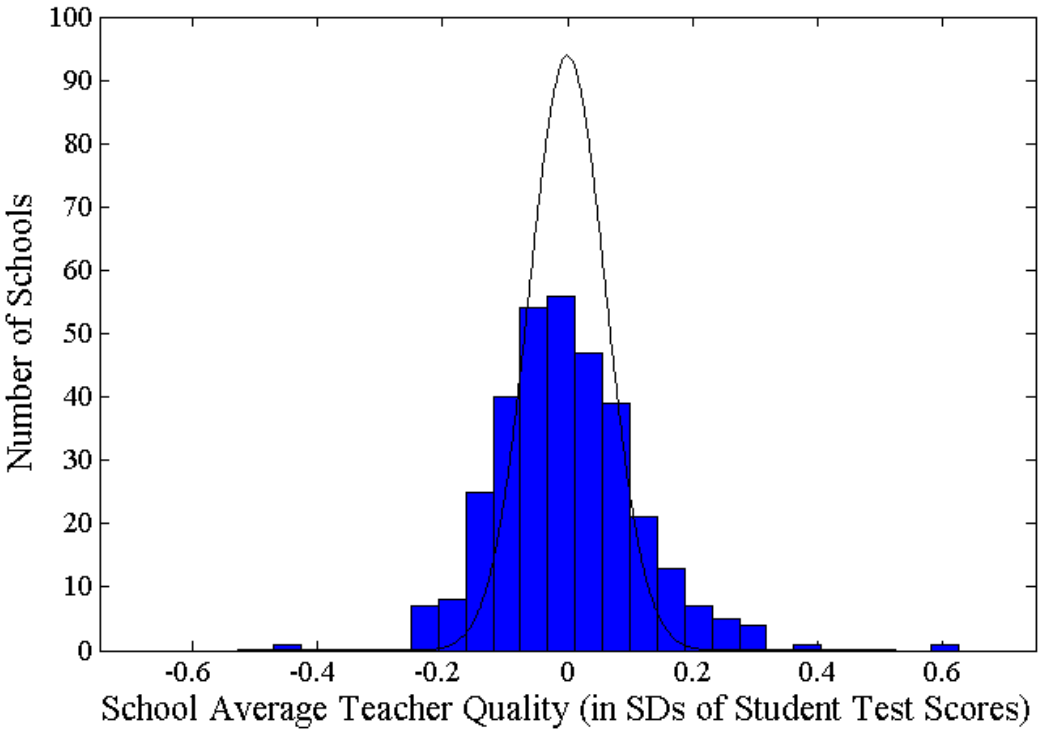


Figure 4: Estimates of School Sensitivity to Teacher Quality ($\hat{\gamma}_s$) and the Underlying Density of School Sensitivity to Teacher Quality (γ_s), after Correcting for Sampling Error

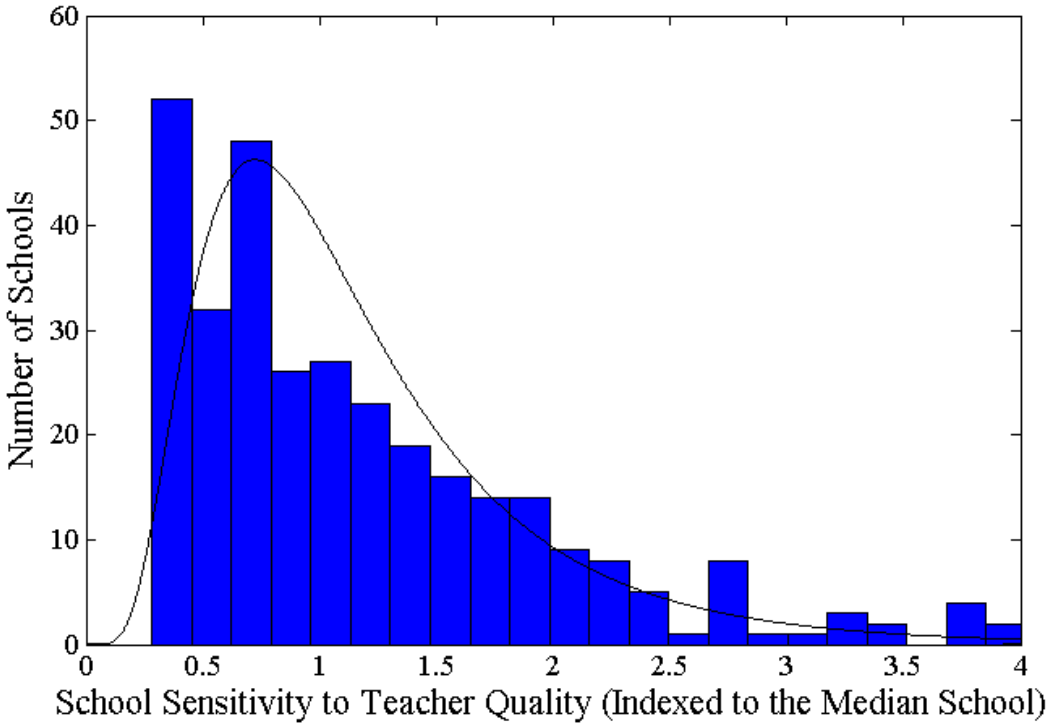


Figure 5: School Quality Estimates ($\hat{\delta}_s$) and the Underlying Density of School Quality (δ_s), after Correcting for Sampling Error

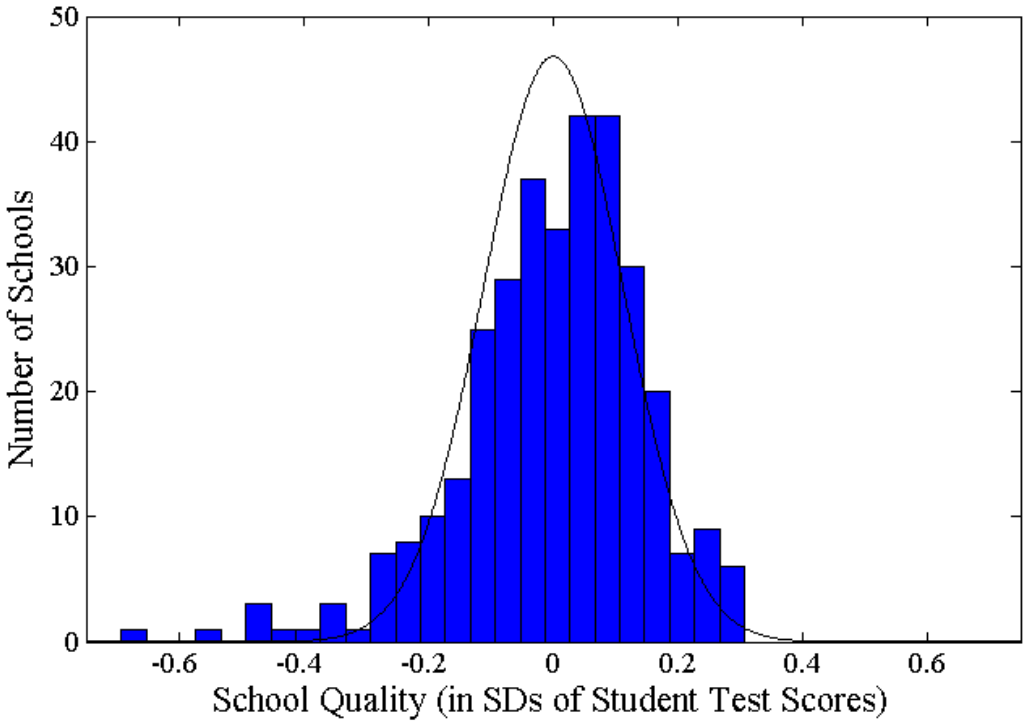
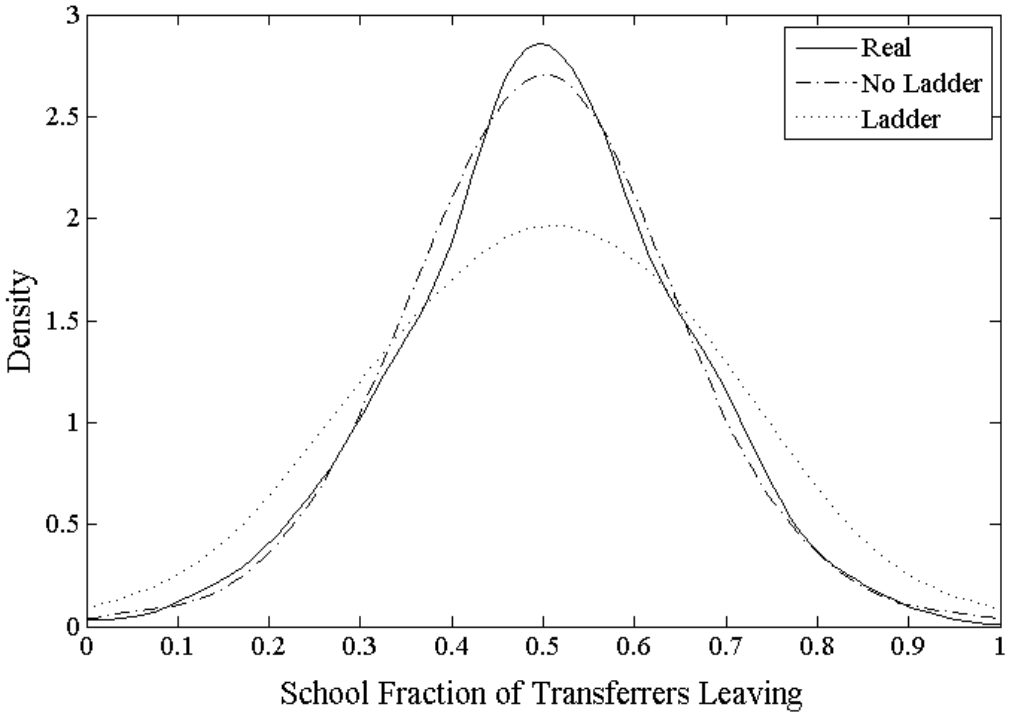


Figure 6: Testing for the Existence of a Job Ladder: A Plot of the Distribution Across Schools of the Fraction of Associated Transferring Teachers That Are Leavers (vs. Arrivers) Using Sample Data, Simulation with No Ladder, and Simulation with Ladder



Appendix

1 Proof of Identification

Proposition:

Consider a set of schools $\mathcal{S}$ and a set of teachers $\mathcal{R}$, each of whom has taught at a school in $\mathcal{S}$. Suppose there exists a subset of teachers, $\tilde{\mathcal{R}} \subset \mathcal{R}$, who have taught at multiple schools in $\mathcal{S}$ in such a way that a connected graph may be formed with the schools in $\mathcal{S}$ as vertices and the transfers of the members of $\tilde{\mathcal{R}}$ as edges. Suppose further that teachers improve or decline over time for some interval of experience ($\exists\ x \neq x'$ such that $d(x) \neq d(x')$), and that there exists a teacher in each school in $\mathcal{S}$ who is observed at both experience levels x and x'. Suppose also that $\gamma_s > 0$ for all $s \in \mathcal{S}$. Finally, suppose that Assumptions 1-3 hold. Then γ_s and δ_s are identified up to scale for all $s \in \mathcal{S}$, μ_r is identified up to scale for all $r \in \mathcal{R}$, and $d(ex)$ is identified up to scale for all levels of experience observed.

Proof:

We assume for this proof that each school has a large number of teachers, each of whom taught a large number students in each of a large number of years. Consider the set of teachers who have taught at school s while at each of two experience levels x and x', which we denote $\mathcal{R}_s^{x,x'}$. Suppose without loss of generality that each teacher is observed at experience levels x and x' at times t and t', respectively. Also, let Z_{ict} represent the component of student i's test score in course c at time t that is unpredictable on the basis of his observable current and prior inputs: $Z_{ict} = Y_{ict} - X_{ict}\beta_c - \tilde{Y}_i^{t-1}\alpha_c$.[52] Then, comparing the expected performance of students taught by members of $\mathcal{R}_s^{x,x'}$ at time t with those taught at time t', we have:

$$\begin{aligned}
&E[Z_{ict}|s(i,t) = S_1, r(i,c,t) \in \mathcal{R}_s^{x,x'}] - E[Z_{ict'}|s(i,t') = S_1, r(i,c,t') \in \mathcal{R}_s^{x,x'}] \\
&= \gamma_1[d(x) - d(x')] + E[\epsilon_{ict}|s(i,t) = S_1, r(i,c,t) \in \mathcal{R}_s^{x,x'}] - E[\epsilon_{ict}|s(i,t') = S_1, r(i,c,t') \in \mathcal{R}_s^{x,x'}] \\
&= \gamma_1[d(x) - d(x')], \qquad (21)
\end{aligned}$$

where the expectation is over the distribution of ϵ_{ict}, and we invoke Assumption 2 in moving from the second to the third line.

[52] Note that β and α can be identified separately from the other parameters of interest using within teacher-year variation.

Consider two schools, S_1 and S_2. If we construct the moment in (21) for both schools using the sets of teachers $\mathcal{R}_{S_1}^{x,x'}$ and $\mathcal{R}_{S_2}^{x,x'}$, respectively, and take the ratio of these moments, we have:

$$\begin{aligned}&\frac{E[Z_{ict}|s(i,t)=S_2, r(i,c,t)\in\mathcal{R}_{S_2}^{x,x'}]-E[Z_{ict'}|s(i,t')=S_2, r(i,c,t')\in\mathcal{R}_{S_2}^{x,x'}]}{E[Z_{ict}|s(i,t)=S_1, r(i,c,t)\in\mathcal{R}_{S_1}^{x,x'}]-E[Z_{ict'}|s(i,t')=S_1, r(i,c,t')\in\mathcal{R}_{S_1}^{x,x'}]}\\&=\frac{\gamma_2[d(x)-d(x')]}{\gamma_1[d(x)-d(x')]}=\frac{\gamma_2}{\gamma_1}\end{aligned}\tag{22}$$

Since $\gamma_1 > 0$ and $d(x) - d(x') \neq 0$ by assumption, if we normalize $\gamma_1 = 1$ we can identify the relative sensitivity of S_2, γ_2. Making such a comparison for each school in $\mathcal{S}$ identifies the distribution of sensitivities $\{\gamma^S\}$, relative to the sensitivity of the normalized school, γ_1.[53]

Note that ϵ_{ict} includes school-year (ϕ_{st}) and teacher-year (ν_{rt}) error components in addition to student-level measurement error and unobserved inputs. Thus, identification of relative school sensitivities ($\{\gamma_s\}$) requires that teachers in $\mathcal{R}_s^{x,x'}$ collectively teach at the school for a large number of years. In practice, this could occur if there are few teachers in $\mathcal{R}_s^{x,x'}$ but each teaches for a large number of years at school s, or if there are many teachers in $\mathcal{R}_s^{x,x'}$, each of whom teaches in a moderate number of years in such a way that the full distribution of school years is represented.

Reconsider equation (21) above. Since the distribution of $\{\gamma^s\}$ is identified by ratios of difference moments as shown above, the levels of these differences now identify $d(x) - d(x')$. Thus, if we normalize the average quality of first year teachers to be 0, so that $d(0) = 0$, then we can identify $d(ex)$ for all observed experience levels by comparing teachers' performance while at these levels to their performance in the first year.[54]

Now, let $\mathcal{X}$ denote the set of observed experience levels, and let $p_r^s(x)$ denote the fraction of teacher r's career at school s that was spent at experience level x. Also, let $\mathcal{T}_r^s$ denote the set of years in which teacher r taught at school s. For a particular teacher r', the expected level of performance for a student taught during the part of her career spent at school s' can be expressed

[53]Note that if $d(x)$ varies over a number of values of x, we do not need to observe a large set of teachers in each school at any particular combination of x and x', just a large set of teachers observed at multiple experience levels, whatever they might be. Comparisons between two schools can take place over the sets of teachers observed at each set of adjacent experience levels, and the relative school sensitivities will be informed by all such moment comparisons.

[54]Once $d(x)$ has been identified for some values of x, comparisons can be made relative to these levels of experience.

as:

$$E[Z_{ict}|s(i,t)=s', r(i,c,t)=r', t\in\mathcal{T}_r^s] = \delta_{s'} + \gamma_{s'}(\sum_{x\in\mathcal{X}} p_{r'}^{s'}(x)d(x) + \mu_{r'})$$
$$+ E[\epsilon_{ict}|s(i,t)=s', r(i,c,t)=r', t\in\mathcal{T}_r^s])$$
$$= \delta_{s'} + \gamma_{s'}(\sum_{x\in\mathcal{X}} p_{r'}^{s'}(x)d(x) + \mu_{r'}) \tag{23}$$

where we have invoked Assumptions 1 and 2 to move from the second to the third line.

Now, let R_{11} and R_{12} be two teachers who each taught at school S_1. Comparing the expected performance of their students over the course of their respective careers at S_1, we obtain:

$$E[Z_{ict}|s(i,t)=S_1, r(i,c,t)=R_{12}, t\in\mathcal{T}_{12}^1] - E[Z_{ict}|s(i,t)=S_1, r(i,c,t)=R_{11}, t\in\mathcal{T}_{11}^1]$$
$$= \gamma_1(\sum_{x\in\mathcal{X}}(p_{12}^1(x) - p_{11}^1(x))d(x) + (\mu_{12}-\mu_{11})) \tag{24}$$

Since teacher quality is only identified relative to other teachers, we normalize $\mu_{11}=0$. Since d(*) and the set $\{\gamma^s\}$ have already been identified, and we observe the fraction of each teacher's career at S_1 spent in each experience cell, this difference moment identifies μ_{12}. Similar comparisons identify the qualities of all non-transferring teachers at school S_1, $\{\mu_k\}, k\in\mathcal{R}_1/\tilde{\mathcal{R}}_1$.

Finally, since we have a connected graph, there must be some teacher at S_1, R_{1j}, who is a member of $\tilde{\mathcal{R}}$, and thus has taught at another school, S_k. Suppose without loss of generality that R_{1j} transferred from S_1 to S_k. The expected level of student performance of R_{1j} during the part of her career spent at school S_1 is given by:

$$E[Z_{ict}|s(i,t)=S_1, r(i,c,t)=R_{1j}, t\in\mathcal{T}_{1j}^1, R_{1j} \text{ transferred from } S_1 \text{ at some } t'>t]$$
$$= \delta_1 + \gamma_1(\sum_{x\in\mathcal{X}} p_{1j}^1(x)d(x) + \mu_{1j})$$
$$+ E[\epsilon_{ict}|s(i,t)=S_1, r(i,c,t)=R_{1j}, t\in\mathcal{T}_{1j}^1, R_{1j} \text{ transferred from } S_1 \text{ at some } t'>t]$$
$$= \delta_1 + \gamma_1(\sum_{x\in\mathcal{X}} p_{1j}^1(x)d(x) + \mu_{1j}) \tag{25}$$

where we have invoked Assumption 3 to move from lines 2 and 3 to line 4. Thus, comparing the performance of R_{11} and R_{1j} as in (24) identifies μ_{1j}.

Similarly, the expected level of student performance of R_{1j} during the part of her career spent

at school S_k is given by:

$$\begin{aligned}
&E[Z_{ict}|s(i,t) = S_k, r(i,c,t) = R_{1j}, t \in \mathcal{T}^k_{1j}, R_{1j} \text{ transferred to } S_k \text{ at some } t' <= t] \\
&= \delta_k + \gamma_k(\sum_{x \in \mathcal{X}} p^k_{1j}(x)d(x) + \mu_{1j}) \\
&+ E[\epsilon_{ict}|s(i,t) = S_k, r(i,c,t) = R_{1j}, t \in \mathcal{T}^k_{1j}, R_{1j} \text{ transferred to } S_k \text{ at some } t' <= t] \\
&= \delta_k + \gamma_k(\sum_{x \in \mathcal{X}} p^k_{1j}(x)d(x) + \mu_{1j}) \qquad (26)
\end{aligned}$$

where we have again invoked Assumption 3 to move from lines 2 and 3 to line 4. Comparing her performance while at S_k with another teacher at the school, R_{k1}, we have:

$$\begin{aligned}
&E[Z_{ict}|s(i,t) = S_k, r(i,c,t) = R_{k1}, t \in \mathcal{T}^k_{k1}] - E[Z_{ict}|s(i,t) = S_k, r(i,c,t) = R_{1j}, t \in \mathcal{T}^k_{1j}] \\
&= \gamma_k(\sum_{x \in \mathcal{X}} (p^k_{k1}(x) - p^k_{1j}(x))d(x) + (\mu_{k1} - \mu_{1j})) \qquad (27)
\end{aligned}$$

Since d(*), μ_{1j}, and γ_k have been identified above, this difference identifies μ_{k1}. Similar comparisons identify μ_{k*} for the other teachers of S_k, including any teacher at school S_k who is member of $\tilde{\mathcal{R}}$. The expected *level* of test scores for any teacher l during her career at school k gives:

$$E[Z_{ict'}|s(i,t') = S_k, r(i,c,t') = R_{kl}] = \delta_k + \gamma_k(\sum_{x \in \mathcal{X}} p^k_{kl}(x)d(x) + \mu_{kl}) \qquad (28)$$

which identifies δ_k under Assumptions 1 and 2. By continuing to move along the connected graph as we have just done, we can identify μ_r and δ_s for any teacher $r \in \mathcal{R}$ and any school $s \in \mathcal{S}$.

2 Details of Normalization

Recall from Appendix 1 that if the identification conditions are satisfied for S networked schools, only $S-1$ γ parameters, $S-1$ δ parameters, $T-1$ μ parameters, and $J-1$ experience cell effects are identified. Thus, a key consideration is how to choose the normalization so as to ensure that the parameters can be interpreted in a meaningful way. To maintain the sparsity of the design matrices, during estimation we normalize γ^e_l to 1, δ^e_l to 0, and $d^e(0)$ to 0, where the "e" superscript indicates the values obtained after estimation before we re-normalize to achieve our desired interpretation. NLLS is choosing γ^e, δ^e, $d^e(ex)$ and μ^e to fit school-teacher-experience level unpredicted means, which, according to the model, are produced by γ, δ, $d(ex)$, and μ.

Thus, the following equation should hold for a teacher j teaching in school l in year t, subject to sampling error:

$$\delta_l + \gamma_l[\mu_j + d(ex_{jt})] = \delta_l^e + \gamma_l^e[\mu_j^e + d^e(ex_{jt})] \tag{29}$$

Consider a second teacher, h, with the same experience, and suppose that both j and h switch from school l to school m at the same time. Comparing the teachers to each other both at school l and then again at school m, we obtain:

$$\gamma_l^e(\mu_j^e - \mu_h^e) = \gamma_l(\mu_j - \mu_h) \tag{30}$$

$$\gamma_m^e(\mu_j^e - \mu_h^e) = \gamma_m(\mu_j - \mu_h) \tag{31}$$

Recalling that γ_l^e has been normalized to 1, if we take the ratio of these two difference equations, we obtain:

$$\gamma_m^e = \frac{\gamma_m}{\gamma_l} \tag{32}$$

Note that this equation could be iteratively substituted when comparisons are made at any two schools, so it holds for all m. Thus, the estimated sensitivity of a given school is actually the sensitivity relative to the sensitivity of the normalized school. Rather than choose an arbitrary school as the standard, we take each estimated sensitivity and divide it by the median of the estimated sensitivities:[55]

$$\hat{\gamma}_m = \frac{\gamma_m^e}{\text{med}_k(\gamma_k^e)} = \frac{\gamma_m/\gamma_l}{\text{med}_k(\gamma_k/\gamma_l)} = \frac{\gamma_m}{\text{med}_k(\gamma_k)} \tag{33}$$

So one can recover the sensitivity of each school relative to the median school (implying that the median of γ_m^e is 1). Next, focus on just teacher j, and choose another year t':

$$\delta_l + \gamma_l[\mu_j + d(ex_{jt'})] = \delta_l^e + \gamma_l^e[\mu_j^e + d^e(ex_{jt'})] \tag{34}$$

Then taking the difference between the year-specific mean unpredicted test scores of teacher j across t and t' gives:

$$\gamma_l[d(ex_{jt} - d(ex_{jt'})] = \gamma_l^e[d^e[ex_{jt}) - d^e(ex_{jt'})] \tag{35}$$

[55]Since the sensitivities are scaling parameters, they are distributed approximately log-normally, so that the mean is considerably larger than the median. Normalizing so that the mean sensitivity is 1 would imply that the clear majority of schools have sensitivity greater than 1.

Let $ex_{jt} = x$ and $ex_{jt'} = 0$ (so that $d^e(ex_{jt'}) = 0$), and recall that $\gamma_l^e = 1$. Then we have:

$$d^e(x) = \gamma_l[d(x) - d(0)] \tag{36}$$

One can iteratively substitute this expression into differences evaluated at other experience levels and other schools to show that this formula is general. So the estimated effect of a given experience cell returned by NLLS is the effect of being in that cell relative to the omitted cell, when at a school with the sensitivity of the omitted school. If we multiply our estimate by the median of $\{\gamma^e\}$, we obtain:

$$\begin{aligned}\hat{d}(x) &= \text{med}_k(\gamma_k^e)d^e(x) \\ &= \text{med}_k(\frac{\gamma_k}{\gamma_l})\gamma_l[d(x) - d(0)] = \text{med}_k(\gamma_k)[d(x) - d(0)]\end{aligned} \tag{37}$$

Thus, one can recover the expected increase in test scores associated with being in a given experience cell, relative to being a first-year teacher, when teaching at a school of median sensitivity.

Next, reconsider equation 29, but evaluated for teacher j when teaching at the normalized school, l, at the normalized level of experience, 0:

$$\delta_l + \gamma_l[\mu_j + d(0)] = \delta_l^e + \gamma_l^e[\mu_j^e + d^e(0)] = \mu_j^e \tag{38}$$

Revisiting equation 31, and substituting for μ_j^e and γ_m^e, we have:

$$\gamma_m^e(\mu_j^e - \mu_h^e) = \frac{\gamma_m}{\gamma_l}(\delta_l + \gamma_l[\mu_j + d(0)] - \mu_h^e) = \gamma_m(\mu_j - \mu_h) \tag{39}$$

Solving for μ_h^e gives:

$$\mu_h^e = \delta_l + \gamma_l[\mu_h + d(0)] \tag{40}$$

By continuing to make such comparisons between teachers along the connected graph of schools, one can verify that this formula holds for any teacher h. If we compare the difference in estimated qualities for any two teachers, we find:

$$\mu_h^e - \mu_j^e = \gamma_l(\mu_h - \mu_j) \tag{41}$$

To eliminate dependence on the choice of normalized school, we follow the procedure used for $\hat{d}(ex)$, and multiply by the median of $\{\gamma^e\}$:

$$\hat{\mu}_j - \hat{\mu}_h = \text{med}_k(\gamma_k^e)(\mu_j^e - \mu_h^e) = \text{med}_k(\frac{\gamma_k}{\gamma_l})\gamma_l(\mu_j - \mu_h) = \text{med}_k(\gamma_k)(\mu_j - \mu_h) \tag{42}$$

Thus, computing the left hand side for each pair of teachers gives the difference in the ability of the two teachers to increase test scores when both are placed in a neutral school context. We can normalize one μ parameter to be 0, use this equation to trace out the entire distribution, then renormalize the distribution to have a zero mean.

Unfortunately, recovering an interpretable version of the δ parameters is not as easy. Consider again equation 29, evaluated again for teacher j at experience 0, but this time while teaching at school m:

$$\delta_m + \gamma_m[\mu_j + d(0)] = \delta_m^e + \gamma_m^e[\mu_j^e + d^e(0)] \tag{43}$$

If we plug in the expressions found above for γ_m^e and μ_j^e, and solve for δ_m^e, we obtain:

$$\delta_m^e = \delta_m - \frac{\gamma_m}{\gamma_l}\delta_l \tag{44}$$

To eliminate dependence on the choice of normalized school, we add the school's estimated teacher sensitivity multiplied by the mean estimated teacher quality ($\gamma_m^e \frac{1}{R}\sum_r \mu_r^e$):

$$\begin{aligned}\hat{\delta}_m &= \delta_m^e + \gamma_m^e \frac{1}{R}\sum_r \mu_r^e = \delta_m - \frac{\gamma_m}{\gamma_l}\delta_l + \frac{\gamma_m}{\gamma_l}\frac{1}{R}\sum_r(\delta_l + \gamma_l[\mu_r + d(0)]) \\ &= \delta_m + \gamma_m((\frac{1}{R}\sum_r \mu_r) + d(0))\end{aligned} \tag{45}$$

Thus, our estimates of the additive school qualities unfortunately also reflect the true sensitivity of the school to a new teacher of average quality. A strange feature of this non-linear model is that seemingly meaningless assumptions about the decompositions of the level of average test scores in the sample into contributions due to average school quality, average school sensitivity to teacher quality, average teacher quality, and average teacher experience drive components of the estimated variance in school quality. It seems bizarre to claim that the average teacher in North Carolina is increasing student test scores by some amount x, but that the average school is decreasing test scores by the same amount x, or vice versa. Schools can only be compared relative to schools, teachers relative to other teachers, and experience levels relative to other experience levels. Furthermore, the test scores used as a dependent variable do not actually have a natural scale (they have all been standardized to have zero mean and unit variance to facilitate comparison across subjects), so the level of the average test score in the sample is meaningless as well. We will assume in interpreting school additive effects that $\frac{1}{R}\sum_r(\mu_r) + d(0) = 0$. Then, differences in estimated school additive effects can be interpreted as differences in the two schools' abilities to increase test scores.

3 Matching Teachers to Students

The NCERDC raw data contains two distinct types of files. The End of Course (EOC) files contain test score level observations for a certain subject in a certain year. Each observation contains various student characteristics, including, importantly, the race, gender, grade level, and gifted status of the student associated with the test score in question. It also contains the class period, course type (which generally indicates academic level), subject code, test date (which generally indicates the semester), school code, and teacher ID code. Unfortunately, the teacher ID corresponds to the teacher who administered the exam, which, particularly in high school, cannot be assumed to be the teacher that taught the class (although in many cases it will be). However, a unique combination of the latter six pieces of information allows me to group students into classrooms. The Student Activity Report (SAR) files contain classroom level observations for a certain year. Each observation contains a teacher ID code (in this case, the actual teacher that taught the class), school code, subject code, academic level, and section number. It also contains the class size, the number of students in each grade level in the classroom, and the number of students in each race-gender cell. Thus, in order to match students to the teacher who taught them, unique classrooms of students in a given subject-school-year combination in the EOC data need to be matched to the appropriate classroom in the SAR data. In small schools, this is often trivial, because there is only one teacher in a given subject in a year, so any physics classroom in the EOC dataset can be safely attributed to the single physics teacher. In large schools, there may be four physics teachers, each teaching four sections, making this process much more subtle.

To overcome this problem, we match the class sizes, grade level totals, and race-gender cell totals of the classrooms in the two datasets. So if one finds exactly one Chemistry class in School 1 in 1999 in both files that has 10 white females and 2 black males, with 5 11th graders and 7 10th graders, one declares a match and removes the classes from the list of classes to be matched. Unfortunately, the SAR data is collected at the beginning of the semester, and the EOC data is collected at the end of the semester. Thus, students who change levels, change sections, or change schools mid-semester will prevent a perfect match from being identified. Thus, we have implemented an iterative fuzzy matching algorithm:

1. Find the absolute difference between each set of matchable classrooms in the following 11 categories: class size, number in each of four grade levels, and number in each of six race-gender cells (hispanic/black/white by male/female).

2. Find pairs of classes that are identical in all 11 categories. If each member of a given pair is only matched identically to its partner in the other dataset (and not a second SAR classroom, for example), the match is made permanent, and these classes are removed from the set of eligible classrooms in the SAR and EOC, respectively.

3. Find remaining pairs of classes that are identical in 10 of the 11 categories. If each member of a given pair only meets this standard with respect to its partner in the other dataset, the match is made permanent, and these classes are removed from the set of eligible classrooms in the SAR and EOC, respectively.

4. Find remaining pairs of classes that are within one unit of each other in all 11 categories. If each member of a given pair only meets this standard with respect to its partner in the other dataset, the match is made permanent, and these classes are removed from the set of eligible classrooms in the SAR and EOC, respectively.

5. Continue lowering the standard in the manner of steps 3) and 4), until there is no pair of remaining classes for which 9 categories are within 5 units of each other. Classrooms that remain are deemed unmatchable, and discarded.

6. If more than one classroom in the SAR dataset is matched to a given classroom in the EOC dataset at a given standard, but the teacher is the same in each of the SAR classrooms, that teacher is matched to the EOC classroom.[56]

7. If two classes do not meet the match standard, but they are the only two remaining classes in the school-subject-year cell, and the teacher id's match, this teacher is matched to the EOC classroom.

8. For those classes that remain unmatched because they meet the exact same standard with multiple classes in the opposing dataset, repeat steps 1-7, except replace differences in grade totals with indicators for whether the course type in the EOC data matches the academic level in the SAR data, and whether the test date in the EOC data matches the semester in the SAR data.

[56]Note that this implies that we do not always know what academic level an EOC classroom was taught at, since we can't always uniquely identify the classroom in the SAR dataset, even if we can uniquely identify the teacher.

9. Repeat steps 1-8, but with percentage differences in each race-gender cell (from the beginning), and percentage differences in each grade level total. This provides a second set of classroom matches.

10. Compare the matches from steps 1-8 with the matches from step 9. If a given classroom is matched to distinct opposing classrooms in the two match algorithms, dissolve the matches. If it is matched to the same opposing classroom in each algorithm, retain the match. If a pair of classrooms are matched in one algorithm, but unmatched in the other, retain the match.[57]

11. Redo 1-10, but decrease the standard more quickly at each iteration. Compare the final matches from this version of the algorithm to the final matches from 10, and dissolve matches where a classroom is matched to different opposing classrooms in the different algorithms. [58]

Frequently, fuzzy matching algorithms like these use a continuous weighting function over the 11 categories to evaluate the quality of the match, and relax the function value iteratively, instead of imposing a strict difference standard for each category, adding up the number of categories that meet this standard, and relaxing this standard iteratively. We chose the latter approach because of its tolerance for typos. Standard weighting functions are usually convex in differences in each category, so that having a large difference in one category severely reduces the quality of the match. However, there were a number of cases in which a classroom in one dataset would have zeros for all the race-gender totals, or an outlandish class size, and we wanted an algorithm that would not punish too much matches which generally fit well, but had one or two categories with large differences. The fraction of classrooms matched varied with the subject, ranging from around 79% for Algebra 1 to 92% for Physics (since fewer people take Physics, there are many fewer sections and teachers, making it much easier to match). If we imposed a strong match standard, in which the algorithm in steps 1-8, 9, and 11 all had to agree on a given pair in order

[57] We hope to dissolve these matches, and re-estimate the model using only classrooms paired in both algorithms as a robustness check.

[58] The reason for this step is that if two different classrooms at a school have very similar makeups, dropouts and transfers may make classroom 1 in the EOC dataset, measured at the beginning of the semester, actually match very slightly better with classroom 2 in the SAR dataset, measured at the end of the semester; in steps 1-10, classroom 1 in the EOC dataset will be incorrectly matched to classroom 2 in the SAR dataset, while in this step, a larger standard drop in a given iteration will mean that classroom 1 in EOC will now meet the same new standard with classroom 1 and classroom 2 in SAR at the same time, and the algorithm will let the semester/academic level information decide which classes get matched, instead of the very subtle difference in the quality of the race-gender distribution match.

for the match to be verified, the fraction of classrooms matched ranged from 50% in Algebra 1 to 85% in Physics.[59]

4 Details of Estimation

If we stack all test scores into a single vector, we can rewrite (9) in matrix form as:

$$\mathbf{Y} = \tilde{\mathbf{Y}}\alpha + \mathbf{X}\beta + \mathbf{C}\delta + \mathbf{C}\gamma[\mathbf{M}\mu + (\mathbf{Ex})\mathbf{d}] + \epsilon \tag{46}$$

where:

$\mathbf{Y}$ is an $Nx1$ vector of standardized test scores, aggregated across classes, courses, schools, and years. Each test score is standardized relative to the distribution of test scores from the relevant subject in the relevant year.

$\tilde{\mathbf{Y}}$ is an NxL matrix of prior (pre-high school) test scores and squares of prior test scores.

$\mathbf{X}$ is an NxK matrix of covariates. Some covariates are at the classroom level, some are at the student level. All covariates are fully interacted with subject indicators. Note that many students have test scores in a number of high school subjects.

$\mathbf{C}$ is an NxS design matrix in which $\mathbf{C}(i,j) = 1$ if test score i is associated with a class taken in school j.

$\mathbf{M}$ is an NxR design matrix in which $\mathbf{M}(i,j) = 1$ if test score i is associated with a class taught by teacher j.

$\mathbf{Ex}$ is an NxJ design matrix in which $\mathbf{Ex}(i,j) = 1$ if test score i is associated with a class in which the teacher was in experience cell j.

$\mathbf{d}$ is a $Jx1$ vector of parameters that indicates how much an average teacher in the corresponding experience cell increases test scores, relative to a first year teacher.

ϵ is an $Nx1$ vector of measurement errors and unobserved inputs.

First, note that while $\mathbf{C}$ and $\mathbf{M}$ are huge matrices, they are extremely sparse, so that employing algorithms designed for sparse matrices considerably reduces the amount of memory required. Second, note that equation 4 can be rewritten in the following way:

$$Y_{ict} = \mathbf{X_{it}}\beta_{\mathbf{c}} + \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}} + \lambda_{ct} + \epsilon_{ict} \tag{47}$$

where

$$\lambda_{ct} = \lambda_{srj} = \delta_{s(i,t)} + \gamma_{s(i,t)}[d(ex_{r(i,c,t)}) + \mu_{r(i,c,t)}] \tag{48}$$

[59]Recall that the weaker standard still does not tolerate conflicts, but does tolerate one of the algorithms failing to match a class at all, as long as the second does.

In other words, one can first think of each test score as a combination of current and past family, individual, and peer inputs, and a school-teacher-experience-specific effect. This suggests a two-stage approach, in which the first stage estimates school-teacher-experience combination effects, and the second stage decomposes these combination effects into additive school effects (δ), school sensitivities (γ), experience profiles $d(ex)$, and teacher effects (μ). The first stage estimates the following equation:

$$Y = \mathbf{X}\beta + \tilde{\mathbf{Y}}\alpha + \mathbf{A}\lambda + \zeta \tag{49}$$

where $\mathbf{A}$ is an NxH matrix, with H denoting the number of observed school-teacher-experience level combinations. $\mathbf{A}(i,j) = 1$ if test score i was achieved in the j-th teacher-school-experience combination. ζ is the component of ϵ that is within school-teacher-experience combinations.

The second stage estimates the following equation:

$$\hat{\lambda} = \tilde{\mathbf{C}}\delta + \tilde{\mathbf{C}}\gamma[\tilde{\mathbf{M}}\mu + \tilde{\mathbf{exD}}] + \omega \tag{50}$$

where $\tilde{C}$ is an HxS matrix such that $\tilde{C}(i,j) = 1$ if school-teacher-experience effect i is associated with school j,

$\tilde{M}$ is an HxR matrix such that $\tilde{M}(i,j) = 1$ if school-teacher-experience effect i is associated with teacher j,

$\tilde{ex}$ is an HxJ matrix such that $\tilde{exp}(i,j) = 1$ if school-teacher-experience effect i is associated with teacher experience cell j, and

$\omega(i)$ is the component of ϵ common to students in school-teacher-experience combination i.

Given that β and α are very precisely estimated using only within teacher-school-experience cell variation, estimating equation 46 using a two-stage approach results in virtually no loss of efficiency relative to the one stage approach. However, this approach has a couple of important computational advantages. First, the first stage is linear, and can thus be estimated by OLS. Abowd et al. (2002) show that by expressing the OLS estimator as $(X'X)B = X'Y$, one can use row-reduction to solve for B without needing to calculate $(X'X)^{-1}$, which would impose a considerable computational burden. The resulting estimates $\hat{\lambda}$ equal the mean test scores associated with a given school-teacher-experience combination, net of the effects of the X covariates and the $\tilde{Y}$ vector of prior test scores:

$$\hat{\lambda}_h = \frac{1}{N_{i \in h}} \sum_{i \in h} (Y_i - X_i\hat{\beta} - \tilde{Y}_i^{t-1}\hat{\alpha}) \tag{51}$$

Second, the second stage, where nonlinear estimation is necessary, now involves $\tilde{D}$ and $\tilde{M}$, which are HxS and HxR instead of NxS and NxR. Third, notice that the identification argument given above relied exclusively on across school-teacher and across-experience cell-within teacher variation. Teachers that are only observed teaching within one experience cell contribute nothing to the identification of δ, $d(exp)$, γ, nor the μ parameters associated with any other teachers. Specifically, the quality of each single experience cell teacher can be chosen to match exactly the mean unpredicted test score associated with that teacher. Thus, first stage means associated with single experience cell teachers can be dropped during second stage estimation, along with the columns in $\tilde{M}$ associated with single experience cell teachers. Once the γ and δ parameters have been estimated, one can then estimate the remaining μ parameters of the single experience cell teachers by choosing μ to fit their mean unpredicted test score. This greatly reduces the number of parameters being estimated, since about 27% of the teachers in my sample are only observed in one experience cell, which makes non-linear least squares computationally feasible.[60]

5 Calculation of Standard Errors

Mimicking the estimation procedure documented in Appendix 4, standard errors are estimated in two stages. First, we calculate the variance of student-level observable coefficients ($\hat{\beta}_c$ and $\hat{\alpha}_c$) and school-teacher-experience cell combinations ($\hat{\lambda}$) using the standard formula for OLS asymptotic variance: $V = (G'G)^{-1}G'\Omega G(G'G)^{-1}$, where in our context $G = [X, \tilde{Y}, A]$ and $\Omega = var(\zeta)$. Then, in the second stage, we apply the standard formula for weighted NLLS asymptotic variance, using the estimated school-teacher-experience effects $\{\hat{\lambda}\}$ as observations: $\Sigma = (J'WJ)^{-1}J'WVWJ(J'WJ)^{-1}$. The weighting matrix W is a diagonal H x H matrix that weighs each estimated school-teacher-experience effect $\hat{\lambda}$ by the number of exam scores in the corresponding school-teacher-experience cell. J is the H x $(2S + R + J)$ Jacobian matrix of partial derivatives of the school-teacher-experience residuals with respect to the parameters $\{\hat{\delta}\}$, $\{\hat{\mu}\}$, $\{\hat{\gamma}\}$ and $\{\hat{d}(ex)\}$. J can be calculated analytically, given the relatively simple non-linear form of the production function.

However, the relative simplicity of these variance formulas belies the considerable computational difficulty associated with their evaluation. Recall that there are $N = 4,016,343$ test-score

[60]Note that this does not imply that 27% of my sample only teach in one experience cell. We only observe teachers when they teach one of ten subjects, so many of the single experience cell teachers are teachers in different subjects who were called upon to teach one of the ten we observe in only one year or time interval.

level observations, $K + L = 800$ subject-specific coefficients on student background characteristics and prior test scores, and $H = 33,153$ school teacher experience cells. Direct evaluation of V would require both the inversion of a $33,953$ x $33,953$ matrix $(G'G)$ and the construction of a 4 million x 4 million matrix (Ω). Both of these operations exceed the memory limits of even very powerful servers. A couple of subtle tricks were necessary to make this calculation feasible within a reasonable length of time. First, note that V can be written as the product $V = AGB$, where A is the H x N matrix $(G'G)^{-1}G'\Omega$, G is N x H, and B is the H x H matrix $(G'G)^{-1}$. Next, let $A(k)$ denote the k-th column of A, and define A^k as the H x N matrix in which the k-th column consists of $A(k)$, and all other elements are zeros. Note that A can be written as:

$$A = A^1 + A^2 + ... + A^N \tag{52}$$

We can calculate $A^k, k = 1, ..., N$, as follows. First, we construct the k-th column of Ω, denoted $\Omega(k)$. Next we solve the linear system $(G'G)A(k) = G'\Omega(k)$ using Cholesky factorization to recover $A(k)$. Then, we create an H x N matrix of zeros, and substitute the k-th column with $A(k)$ to obtain A^k. Since only the k-th column of A^k has non-zero entries, we can store A^k easily in memory as a sparse matrix. Breaking A up into these N distinct pieces facilitates the use of parallel processing. This prevents statistical software from running out of working memory on any given processor, and speeds up computation considerably.

While this procedure allows us to avoid both calculating $(G'G)^{-1}$ directly and constructing Ω, we cannot simply sum $A^1, ..., A^N$ to recover A; A is still H x N, which is too large to load into working memory on a single processor. We overcome this problem by post-multiplying each A^k by G before summing, leaving the H x H matrix AG:

$$AG = A^1G + A^2G + ... + A^NG \tag{53}$$

While post-multiplying by G removes sparsity, such sparsity is no longer necessary, since AG is only H x H. Finally, in order to avoid calculating $(G'G)^{-1}$ directly, we calculate V row-by-row by solving the linear system $V'(k)(G'G) = (AG)'(k)$. We concatenate the $V'(k)$ and transpose to recover V. An analogous procedure is employed to recover Σ, with V taking the place of Ω, and JW taking the place of G.

6 Proof of Identification with Endogenous Mobility toward Better Match Quality, when Mobility is Balanced

This section amends the identification proof in Appendix 1 for the case in which teachers' transfer decisions may be based in part on their current or potential match quality, κ_{rs}. In place of Assumption 3, we assume instead that a large set of transfers connects each school, and that mobility is balanced at each school, so that the number of transfers into each school equals the number of transfers out. We also assume that the strength of movement driven by match quality does not depend on the school: $E[\kappa_{rs}|r \text{ transferred from } s] = c$ and $E[\kappa_{rs}|r \text{ transferred to } s] = d$, for some constants c and d, for all schools s. This proof mirrors its analog from Appendix 1 until the identification of the quality of transferring teachers, beginning in the paragraph before equation (25). Hence, the set $\{\gamma_s\}$, the function $d(*)$, and the qualities of non-transferring teachers at school S_1, $\{\mu_r|r \in \mathcal{R}_1/\tilde{\mathcal{R}}_1\}$ are identified as before. We show that the set of school qualities, $\{\delta_s\}$, and the set of non-transferring teacher qualities, $\{\mu_r|r \in \mathcal{R}/\tilde{\mathcal{R}}\}$, are still identified.

Let $\tilde{\mathcal{R}}_s^o$ denote the (large) set of transferring teachers who taught at school s and then transferred out to a different school, and let $p^s_{\tilde{\mathcal{R}}^o}(x)$ denote the fraction of total years spent at s among members of $\tilde{\mathcal{R}}_s^o$ that were spent at experience level x. Similarly, let $\tilde{\mathcal{R}}_s^i$ denote the set of transferring teachers who transferred in to s, and define $p^s_{\tilde{\mathcal{R}}^i}(x)$ analogously. If mobility is balanced, then the expected residual among transferring teachers associated with school S_1 while at school S_1 is an equally weighted average of two components:

$$
\begin{aligned}
&E[Z_{ict}|s(i,t) = S_1, r(i,c,t) \in \tilde{\mathcal{R}}_1] \\
&= \frac{1}{2}(E[Z_{ict}|s(i,t) = S_1, r(i,c,t) \in \tilde{\mathcal{R}}_1^o) + \frac{1}{2}(E[Z_{ict}|s(i,t) = S_1, r(i,c,t) \in \tilde{\mathcal{R}}_1^i]) \\
&= \frac{1}{2}(\delta_1 + \gamma_1(\sum_{x \in \mathcal{X}} p^1_{\tilde{\mathcal{R}}^o}(x)d(x) + E[\mu_r + \kappa_{r1}|r \in \tilde{\mathcal{R}}_1^o]) \\
&+ \frac{1}{2}(\delta_1 + \gamma_1(\sum_{x \in \mathcal{X}} p^1_{\tilde{\mathcal{R}}^i}(x)d(x) + E[\mu_r + \kappa_{r1}|r \in \tilde{\mathcal{R}}_1^i])) \qquad (54)
\end{aligned}
$$

where the expectation in this case is over the choice of teacher within the set of transferring teachers (with the probability that a given teacher is chosen given by the fraction of students collectively taught by the set at that school that were taught by the particular teacher), as well as over the distribution of ϵ_{ict}. We have also invoked Assumption 1 and 2 to move from line 1 to line 2. Recall that γ_1 has been normalized to 1, δ_1 has been normalized to 0, and $d(*)$ can be

identified using within-school variation in performance across experience levels. Thus, comparing this moment with the moment in equation (25), we can see that in the absence of endogenous mobility, this moment would identify the true average teaching quality of transferring teachers. Instead, with movement driven by match quality, we identify a combination of the average teaching quality and the average match quality of transferring teachers at the school, averaged across transferrers that left and came: $E[\mu_r + \kappa_{r1} | r \in \tilde{\mathcal{R}}_1]$. To the extent that transferring teachers are on average less well matched over the course of their careers, their qualities will be underestimated relative to non-transferring teachers.

Now, suppose that all members of $\tilde{\mathcal{R}}_1$ transferred to or from school S_k (so $\tilde{\mathcal{R}}_1 = \tilde{\mathcal{R}}_k$ and mobility at S_k is also balanced).[61] Then the expected average residual among transferring teachers associated with S_k is given by:

$$
\begin{aligned}
&E[Z_{ict} | s(i,t) = S_k, r(i,c,t) \in \tilde{\mathcal{R}}_1] \\
&= \frac{1}{2}(E[Z_{ict} | s(i,t) = S_k, r(i,c,t) \in \tilde{\mathcal{R}}_1^o]) \\
&+ \frac{1}{2}(E[Z_{ict} | s(i,t) = S_k, r(i,c,t) =\in \tilde{\mathcal{R}}_k^i]) \\
&= \frac{1}{2}(\delta_k + \gamma_k(\sum_{x \in \mathcal{X}} p^k_{\tilde{\mathcal{R}}^o}(x) d(x) + E[\mu_r + \kappa_{rk} | r(i,c,t) \in \tilde{\mathcal{R}}_1^i])) \\
&+ \frac{1}{2}(\delta_k + \gamma_k(\sum_{x \in \mathcal{X}} p^k_{\tilde{\mathcal{R}}^i}(x) d(x) + E[\mu_r + \kappa_{rk} | r(i,c,t) \in \tilde{\mathcal{R}}_1^o])) \qquad (55)
\end{aligned}
$$

Note that $d(*)$ and $\{\gamma_s\}$ have already been identified from variation in performance across experience levels. Furthermore, $E[\mu_r + \kappa_{rk} | r \in \tilde{\mathcal{R}}_1] = E[\mu_r + \kappa_{r1} | r \in \tilde{\mathcal{R}}_1]$ by the strength of mobility assumption stated above, and $E[\mu_r + \kappa_{r1} | r \in \tilde{\mathcal{R}}_1]$ was identified by the moment in equation 55. Thus, this moment identifies δ_2. The level of performance of each non-transferring teacher at school k then identifies her quality, as in the proof of the baseline model. Thus, mobility driven by match quality undermines the identification of the relative quality of transferring teachers compared to non-transferring teachers, but if mobility is balanced, it does not undermine identification of relative school quality, nor the identification of school sensitivity to teacher quality or relative average quality of non-transferring teachers among different schools. Overall average teaching quality at a school will only be biased to the extent that the school has a relatively

[61] I am almost certain that this is actually without loss of generality, but the proof requires a very different form if mobility is not balanced across all pairs of schools (but is still balanced overall for each school). Rather than identifying parameters sequentially as we have done here, we would need to show that the system of moment equations relating average student residuals in each school-teacher-experience level cell to the underlying parameters has a unique solution.

large or small fraction of its teachers that are transferrers, and the average match quality of transferrers over their careers is different from that of non-transferring teachers.

7 Appendix Figures

Figure 1: Distributions of Standardized Scores by Subject-Year: Part 1

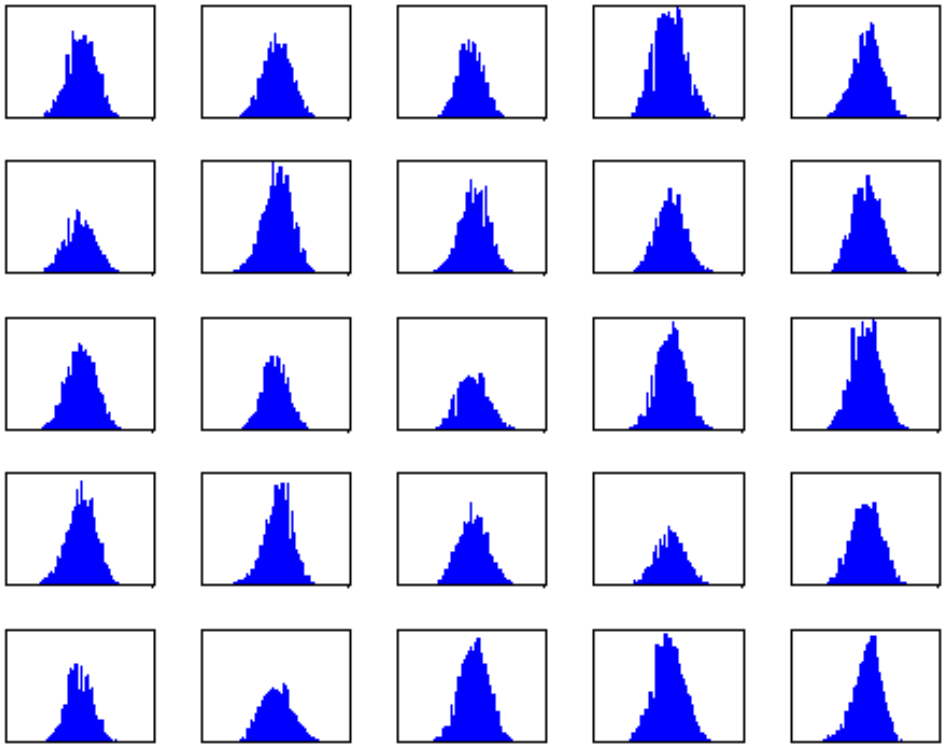


Figure 2: Distributions of Standardized Scores by Subject-Year: Part 2

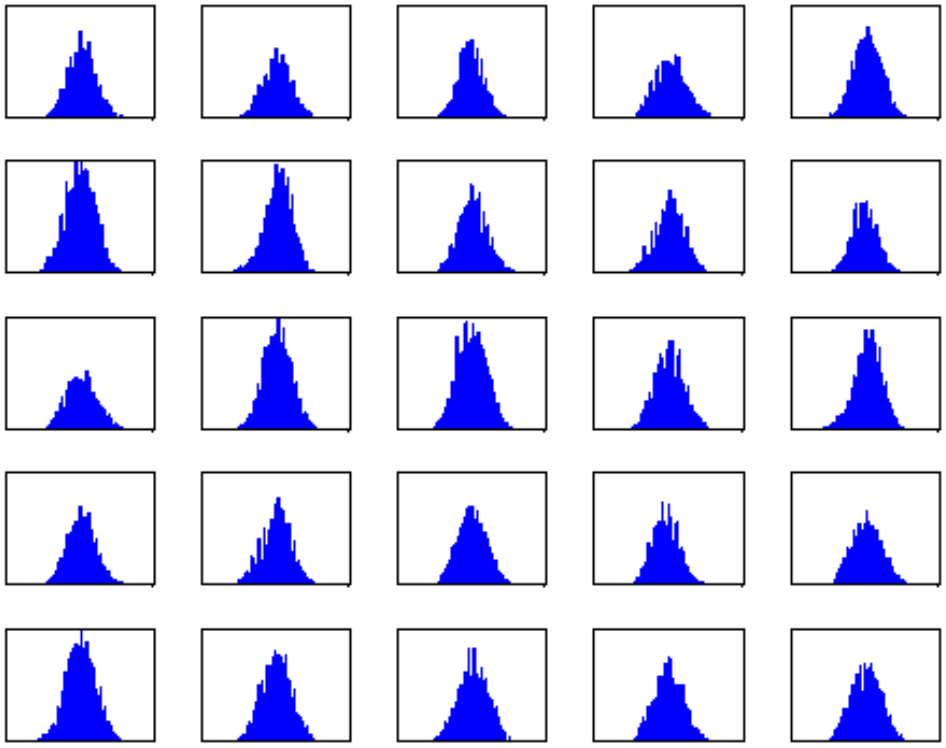


Figure 3: Distributions of Standardized Scores by Subject-Year: Part 3