Teacher Quality and Student Inequality [*]

Rick Mansfield

September 22, 2010

1 Introduction

Recent research using matched student-teacher data has confirmed the existence of an important role for teaching quality in producing student test score improvement in elementary and middle schools (Aaronson, Barrow, Sander (2007), Hanushek, Kain, O'Brien, Rivken (2005), Kane, Rockoff, and Staiger (2007)). This development 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 certain kinds of schools, policies that incentivize relocation of high quality teachers may not succeed in trimming student performance gaps across schools if the targeted schools are relatively insensitive to teacher quality. For example, some schools' facilities may be in such a state of disrepair, or their students may be so unruly or ill-prepared, that even good teachers cannot manage to raise their performance. Thus, this paper aims to answer two questions. First, how equitably is teaching quality distributed within and across public high schools? And second, would the performance of students at underperforming schools improve substantially if they were taught by better teachers?

To address these questions, we exploit administrative data from the North Carolina

[*] PRELIMINARY AND INCOMPLETE. DO NOT CIRCULATE WITHOUT AUTHOR'S PERMISSION. 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.

Education Research Data Center that permits high school students in the universe of North Carolina public high schools to be matched to their teachers and test scores in up to ten high school courses from 1997-2006. Such rich data 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 interact with a teacher's quality. 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 such complementarity between school and teacher quality creates scope for efficiency gains from incentive schemes that reallocate teachers. Separate identification of teacher quality from two dimensions of school quality stems from a large network of teacher transfers, coupled with a testable exogenous mobility assumption.

To this point, nearly all of the attempts to recover within-school teacher fixed effects involve elementary or middle school test scores.[1] This may be largely attributed to data inavailability; in many states, standardized exams at the high school level are either not mandated or are not closely linked to course curricula. One might also argue, though, that since high school teachers only interact with each class of students for about one hour each day, they are likely to have smaller effects on student test scores. Furthermore, if one believes that the true education production function involves complementarities between current and past inputs, then high school may already be too late to help students who have not received proper support in earlier years. However, there are a number of distinct advantages to using high school data. 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 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, and with it the possibility that the worst schools with the least-supported students are also receiving inferior teaching. Third, high school teachers often teach four or five different classrooms each year, so that teachers may teach over 100 students a year, and over 1000 students in 10 years. Hence, although

[1] These include Aaronson et al (2007), Boyd et al (2007), Hanushek, Rivken, and Kain (2005), Kane and Staiger (2008), Kane et al. (2007), Goldhaber(2007) and Rothstein(2010).

the effect sizes may be smaller, they can be more precisely estimated.

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 scores 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, only about 1% 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. Furthermore, variation in teacher experience across schools contributes almost nothing to across-school test score gaps.

Schools do seem to have 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 reallocate teachers 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 average test scores by .09 student level standard deviations *and* decrease test score variance by about 5%.

The rest of the paper is structured as follows. Section 2 presents the model to be estimated, which consists of an educational achievement production function along with the assumptions required to justify its simple form. Section 3 discusses identification of the model. Section 4 presents the estimation strategy. Section 5 discusses the treatment of measurement error in test scores. Section 6 describes the data. Section 7 presents the estimated distributions of teacher quality and school sensitivity to teacher quality, followed by a detailed look at which kinds of schools seem to be attracting the best teachers, and which kinds are most sensitive to teacher quality. In Section 8 we use these estimated distributions to examine the impact of three 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. Second, we examine the potential impact of an intervention proposed in the popular media: using noisy estimates of teacher quality based on student test score performance to make teacher tenure decisions. Briefly, we find that the predicted efficacy and fairness of such a policy depends critically on the chosen specification of the education production function. Finally, given the sizable within-school variation in teacher quality, we also examine the extent to which the average quality of teaching received over a high school career varies across students, and how this student-level variation in teacher quality would change if students were randomly assigned to teachers within schools. Section 9 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 do not reveal any discernable job ladder among public high schools. Section 10 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 students $\{\mathbf{A_i^t}\}$, 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.[2] We also assume that current teacher and school inputs are additively separable from current student, family, and peer inputs, and from the measurement error.[3] We obtain:

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

Let the school and teacher associated with student i at year t be denoted by $s(i,t)$ and $r(i,t)$, respectively. We assume that school and teacher inputs enter the production function as a

[2] 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 topics, 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). 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.

[3] 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 limited 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.

linear combination of current school quality $q_{s(i,t)}$ and current teacher quality $q_{r(i,t)}$:

$$h(\mathbf{R_{it}}, \mathbf{S_{it}}) = q_{s(i,t)} + q_{r(i,t)} \tag{2}$$

We decompose current school quality into a permanent component, $\delta_{s(i,t)}$, and a transient component, $\omega_{s(i,t)t}$. The $\delta_{s(i,t)}$ parameters capture all school level conditions that affect student learning independently of teacher quality. These include principal quality, safety of the neighborhood, and quality of school facilities. The transient component $\phi_{s(i,t)t}$ captures fluctuations in principal quality, crime waves, captures renovations of school facilities, etc.

Current teacher quality, on the other hand, is assumed to reflect five components. 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 increase test scores, $\mu_{r(i,t)}$.

Second, other research indicates that a teacher's effectiveness increases substantially with at least the first few years of experience (Clotfelter et. al. (2007)). Thus, $d(ex_{r(i,t)t})$ captures predictable growth in teacher effectiveness with experience, $ex_{r(i,t)t}$, which is assumed to be common across teachers.

Third, idiosyncratic 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.

Fourth, $\kappa_{r(i,t)s(i,t)}$ 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 correspond coincide with how the principal wants lesson plans to be organized, or classrooms to be managed.

Finally, the extent to which a school's students respond to the quality of teaching they receive is captured by a school-specific scaling factor, $\gamma_{s(i,t)}$. The intuition behind the school sensitivity parameter is that some schools may offer such a poor learning environment (due to disruptive students, crowded classrooms, or inadequate supplies) that even a very good teacher cannot raise test scores substantially; no one is learning regardless of who teaches them. 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 sets up the workstations.[4] Thus, we obtain:

$$\begin{aligned} h(\mathbf{R_{it}}, \mathbf{S_{it}}) &= q_{s(i,t)} + q_{r(i,t)} \\ &= \delta_s + \phi_{s(i,t)t} + \gamma_{s(i,t)}[\mu_{r(i,t)} + d(ex_{r(i,t)}) + \nu_{r(i,t),t} + \kappa_{r(i,t),s(i,t)}] \qquad (3) \end{aligned}$$

This specification for the impact of school and teacher inputs contains a number of desirable features. First, the joint distribution of unobservable persistent teacher quality and two dimensions of unobservable persistent school quality is left unrestricted. Second, while persistent teaching quality is assumed to be one-dimensional, the ability of a teacher to raise test scores is nonetheless allowed to vary across schools, as well as over time, as experience accumulates. While one may argue that teachers possess differing quantities of a vector of skills whose relative importance may vary with student or school attributes, estimating multiple dimensions of teacher quality imposes a considerable burden on the limited number of observations per teacher, and defining distinct dimensions of skill is tricky. More importantly, most parents and administrators want to know whether their school has generally "good" teachers, and would be less interested in a decomposition of the schools' teachers into a myriad components. 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.

This specification permits an interpretation of μ_r as the quality of teacher r, and of the distribution of $\{\mu_r\}$ as the distribution of teacher quality.[5] Since the distribution of $\{\gamma_s\}$ will be normalized to have a median of one, μ_r can be thought of as the amount by which a student at a school with median sensitivity to teacher quality can expect his standardized test score to increase (or decrease) by virtue of being assigned teacher r, all else equal. In turn, γ_s can be interpreted as the sensitivity of school s to teacher quality, and its distribution ($\{\gamma_s\}$) will be referred to as the distribution of school sensitivity. $\{\delta_s\}$

[4] 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.

[5] to simplify notation, henceforth $\mu_{r(i,c,t)} = \mu_r$, $\gamma_{s(i,t)} = \gamma_s$, and $\delta_{s(i,t)} = \delta_s$.

will be referred to as the distribution of school quality.

Putting it all together, we have:

$$\begin{aligned} Y_{it} = {} & f(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) + \\ & \delta_s + \phi_{st} + \gamma_s[\mu_r + d(ex_r) + \nu_{rt} + \kappa_{rs}] + e_{it} \end{aligned} \tag{4}$$

Finally, the data we use provides scores from tests based on ten distinct high school subjects, multiple of which may be taken in the same school year. Thus, the education production function in (4) 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 test-score improvement to be specific to the subject being tested, so that $f(*)$ becomes $f^c(*)$ and $g(*)$ becomes $g^c(*)$. However, school quality and teacher quality inputs are assumed to be common to all subjects. In the case of school quality, 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, taken literally, this assumption states that a math teacher must be equally effective at teaching both Algebra 1 and English, but in practice, we only really need to assume that he/she 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 a course-year-specific state distribution, and re-interpret Y_{ict} accordingly. Thus, we have:

$$\begin{aligned} Y_{ict} = {} & f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) + \\ & \delta_s + \phi_{st} + \gamma_s[\mu_r + d(ex_r) + \nu_{rt} + \kappa_{rs}] + e_{ict} \end{aligned} \tag{5}$$

Even with the above assumptions, estimating this 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 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. Thus, the specification we estimate is:

$$Y_{ict} = \mathbf{\tilde{Y}_i^{t-1}}\alpha_\mathbf{c} + \mathbf{X_{ict}}\beta_\mathbf{c} + \delta_s + \gamma_s[d(ex_r) + \mu_r] + \epsilon_{ict} \tag{6}$$

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

$$\begin{aligned}\epsilon_{ict} = {} & (f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) - \mathbf{\tilde{Y}_i^{t-1}}\alpha_\mathbf{c}) + (g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}}) - \mathbf{X_{ict}}\beta_\mathbf{c}) \\ & + \phi_{st} + \gamma_s(\nu_{rt} + \kappa_{rs}) + e_{ict}\end{aligned} \tag{7}$$

While this specification of the education production function still may not be as general as one might like, many richer specifications, such as multiple teacher quality dimensions (e.g. teaching content vs. keeping order) or student-specific sensitivity to teacher quality, are either not identified in the data, or require teachers to teach a massive number of students, so that their qualities could only be reasonably precisely estimated when they are ready to retire.

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 the identification and consistent estimation of an education production function. Thus, to fix ideas and ease discussion of potential biases from estimation, we propose the following timing of input choices. (1) Given the history of school, student, and family inputs up through grade 8, parents choose high schools for their children. (2) Given the set of teachers available to schools in a given year, 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 prior student inputs:

Assumption 1:

$$E[f(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}})|1(s(i,t) = s'); \mathbf{Y}_i^{t-1}, \mathbf{X}_{ict}] =$$
$$E[f(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}})|\mathbf{Y}_i^{t-1}, \mathbf{X}_{ict}] \ \forall \ s' \in \mathcal{S} \tag{8}$$

Then excluding prior student inputs will not bias estimates of persistent school quality (δ_s).[6] In practice, this condition seems unlikely to hold exactly. 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.

The second potential endogeneity problem is that students are not randomly assigned to teachers within schools, so that the average test scores of a given teacher's students may partly reflect deviations in student inputs from school averages. Rothstein (2010), in particular, has revealed the severity of this problem at lower levels of schooling. In order to recover unbiased estimates of persistent teacher quality, μ_r, we need the identity of a student's teacher to give 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:

Assumption 2:

$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|1(r(i,t) = r'), 1(s(i,t) = s'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] =$$
$$E[f^c(\mathbf{F_i^{t-1}}, \mathbf{A_i^{t-1}}, \mathbf{P_i^{t-1}}, \mathbf{R_i^{t-1}}, \mathbf{S_i^{t-1}}) + g^c(\mathbf{F_{it}}, \mathbf{A_{it}}, \mathbf{P_{it}})|1(s(i,t) = s'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}]$$
$$\forall \ r' \in \mathcal{R}, s' \in \mathcal{S} \tag{9}$$

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

[6] 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.

to take an honors class based on private information about how hard they were working in the past or will work in the future. Thus, the validity of this condition 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 6, 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. If these measures serve as an effective proxy for the difficulty level of a class, then conditional random assignment of students to teachers may be a reasonable assumption at the high school level.

However, a third potential endogeneity problem, perhaps more of a concern in this context, is also ruled out by the condition in equation (9). 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 would be underestimated.[7] However, in the non-linear model analyzed here, 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 of the education production function 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 average level of input compensation across all high schools may still bias estimates of the total variance in teacher quality. However, we may also have less reason to be concerned that parental inputs are being chosen to compensate for teacher quality at the high school level. For example, if a first grade teacher fails to teach his/her

[7] 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.

children to read, the children's parents can likely 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 can choose and adjust the quality of teaching they provide. In particular, teachers may be actively choosing their effort level and the content of their lessons. 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 κ_{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, certain aspects of the context surrounding our data help allay these fears. 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 students year-end grade in a given subject, so that parents are likely to complain about teachers that ignore the standard curriculum. Finally, during the sample period in North Carolina, teacher bonuses of up to $1,500 were linked to average test scores of the students in the school at which they teach. Thus, teachers are under considerable pressure to teach the tested material.

A number of existing articles have already shown that the assumptions like those made above are enough to identify and consistently estimate distributions of within-school teacher effects.[8] Comparing the quality of teaching across schools, however, requires extra assumptions and information. Kramarz, Machin and Ouazad (2008) attempt to 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

[8] e.g. Aaronson et al (2007), Kane and Staiger (2008)

inputs, so that the expected change in the test scores of transferring students post-transfer is captured by the difference in the school effects. The second is that schools and student transfers form a connected graph (with schools as vertices and transferring students as edges), so that any two schools may be linked via some chain of student transfers. Consider a special case of the model above in which $\gamma_s = 1 \; \forall \; s$. Then the model in (6) collapses to an alternative version of Kramarz et al (2008) in which teachers take the place of students. Hence, identification requires that (1) schools and teacher transfers form a connected graph, and (2) that a teacher's choice to transfer to a different school does not predict changes either in student and family inputs or in deviations in school or teacher performance from their long-run averages. More formally:

Condition 1: Sufficient Mobility

Consider a set $\mathcal{S}$ of S schools with a set $\mathcal{R}$ of R teachers, each of whom has taught at a school in $\mathcal{S}$. There must exist a subset of teachers, $\tilde{\mathcal{R}} \subset \mathcal{R}$, who teach 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. In other words, between any two schools in $\mathcal{S}$, there must exist a sequence of schools in $\mathcal{S}$ starting and ending with the two selected schools such that from each of the schools in the sequence there is a teacher in $\tilde{\mathcal{R}}$ who transferred between that school and the next school in the sequence.

The sufficient mobility condition is satisfied (easily) in the data. Discussion of the connectedness of the schools in my data takes place in Section 6.

Assumption 3: Exogenous Mobility

$$\begin{aligned} &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\; r' \text{ transferred to } s' \text{ at some } t' <= t] = 0 \\ &E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\; r' \text{ transferred from } s' \text{ at some } t' > t] = 0 \qquad (10) \end{aligned}$$

where ϵ_{ict} is the composite error term defined in equation 7. Given its importance, we will revisit the exogenous mobility assumption and the implications of its violation in Section 9. Essentially, while there are several plausible mechanisms by which the exogenous mobility assumption could be violated, a test of the assumption fails to reject the exogeneity of mobility. Moreover, further analysis indicates that the observed pattern of targeted mobility, in which there is no discernable job ladder, is particularly unlikely to bias the results.

Intuitively, a teacher's persistent quality, μ_r, is 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. 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 need to assume that his/her ability to increase test scores is the same across the two schools in order to compare the average persistent quality of the teachers at these schools. However, with many linking teachers, we need only assume that these teachers are not *systematically* more effective at one of the schools, relative to the non-transferring teachers at that school.

Surprisingly, identification of the non-linear model in equation (6), which includes school-specific sensitivity to teacher quality, does not require much more information. We have defined a teacher's quality as a combination of the average quality of teachers with the same level of experience plus a teacher-specific deviation that is constant over time. 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. A formal proof of identification is provided in Appendix 1. One may be concerned, however, 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 $d_s(ex)$. While this is indeed a concern, 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. In fact, we 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.[9]

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,[10] 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.

Finally, 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 novice teachers have no impact on test scores: $d(0) = 0$. 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 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.

[9] While numerous examples have led me to believe this claim, we have not yet managed to prove it. However, we have proved a similar claim requiring a slightly richer network of transferrers, a requirement that is easily met in the data.

[10] 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.

Given identification, one can shift focus to the precision with which parameters can be estimated. If there are few transferring teachers, but these transferring teachers have taught a large number of students at each school, then their sample school-teacher mean residuals will closely approximate the expected school-teacher mean residuals, and both the teacher effects for these teachers and the relative school sensitivities to teacher quality will be precisely estimated. Alternatively, if each transferring teacher only teaches a few students at each school, then there will still be considerable noise in their sample school-teacher mean residuals from measurement error and idiosyncratic deviations in unobservable inputs from school averages. However, with many transferring teachers, relative school sensitivity can still be estimated precisely, since the average error in the estimated teacher effects among transferrers will tend to zero as the number of transferring teachers gets large (under the assumptions of the model).

In practice, some schools will be connected to the network by more transferrers than others, and will have more teacher-years with which to examine their experience profiles, so that their sensitivity parameters (γ_s) will be more precisely estimated than others. Also, if there are enough students taught by transferrers (either because there are many transferrers or the transferrers taught a large number of students at each school), then the difference in teacher quality for two teachers at different schools that each taught a large number of students will be precisely estimated. In other words, even if we fail to estimate the teacher effects of some of the transferrers very precisely, as long as there are enough transferrers, qualities of non-transferring teachers who taught a lot of students are still precisely estimated, and can be compared across schools.

4 Estimation

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. From section 2, the model to be estimated is:

$$Y_{ict} = \tilde{\mathbf{Y}}_{\mathbf{i}}^{\mathbf{t-1}}\alpha_{\mathbf{c}} + \mathbf{X}_{\mathbf{it}}\beta_{\mathbf{c}} + \delta_s + \gamma_s[d(ex_r) + \mu_r] + \epsilon_{ict} \qquad (11)$$

We can stack the test scores into a single vector, and rewrite (11) in matrix form as:

$$\mathbf{Y} = \tilde{\mathbf{Y}}\alpha + \mathbf{X}\beta + \mathbf{C}\delta + \mathbf{C}\gamma[\mathbf{M}\mu + (\mathbf{Ex})\mathbf{d}] + \epsilon \tag{12}$$

where:

$\mathbf{Y}$ is a 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.

$\mathbf{C}$ 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.

We estimate equation (12) via non-linear least squares. Since students' test scores partly reflect unobserved inputs and measurement error, they are noisy measures of teacher and school quality. Thus, even if the conditional random assignment assumption holds, 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 variation in teacher quality within and across schools. Thus, minimizing N squared deviations over $K + L + 2S + R + J$ parameters is a daunting computational task. Fortunately, there are a few shortcuts that ease the computational burden considerably. Appendix 3 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.

5 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 first define each estimated teacher fixed effect $\hat{\mu}_r$ as the sum of the teacher's true quality and an error component: $\hat{\mu}_r = \mu_r + \xi_r$. Then the sample variance in estimated teacher quality can be decomposed as:[11]

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

Under the assumptions of the model, the error component and a teacher's true quality are uncorrelated. In this case, $2\frac{1}{R}\sum_r(\mu_r\xi_r) \approx 0$.

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

$$V\hat{a}r(\mu_r) = \frac{1}{R}\sum_r(\hat{\mu}_r)^2 - \frac{1}{R}\sum_r(\xi_r)^2 \tag{14}$$

But ξ_r is not observed. However,

$$\frac{1}{R}\sum_r(\xi_r)^2 \approx \frac{1}{R}\sum_r E[(\xi_r)^2] = \frac{1}{R}\sum_r(sd(\xi_r))^2, \tag{15}$$

[11] In theory, one could imagine the set of teachers in my data as a sample from the set of possible high school teachers in North Carolina. Then, the variance in true teaching quality would need to be adjusted to reflect sampling error in the set of teachers observed in my sample. However, the distribution of the quality of potential teachers will vary with state labor market conditions, so that it is not easy to define this population. Thus, we instead consider 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. Furthermore, given the large sample of teachers, the estimates presented will also approximate closely the variance in the quality of possible high school teachers.

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

$$\hat{Var}(\mu_r) = \frac{1}{R}\sum_r (\hat{\mu}_r)^2 - \frac{1}{R}\sum_r (\hat{sd}(\xi_r))^2 \tag{16}$$

We implement the same technique to calculate the true variance in δ_s, γ_s, and school average teacher quality $\overline{\mu}_s$.[12] 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. [13]

6 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. The sample includes all students from all public high schools in North Carolina. During the sample period, North Carolina offered a standard curriculum with mandatory end-of-course tests for the following subjects: Biology, English 1, U.S. History, Econ/Law/Politics, and Algebra 1, Algebra 2, Geometry, Physics, Physical Science, and Chemistry.[14]

The set of observable current student and peer inputs, $\mathbf{X_{ict}}$, includes the student's race and gender, indicators for learning disabilities in writing, math, and reading, limited English proficiency, grade level (9-12), participation in a sport, vocational club, academic club, service club, arts club, whether the student is gifted in math or English, whether the

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

[13] For teacher quality, the formula is:

$$\begin{aligned} \mu_j^B &= \hat{\mu}_j(\hat{Var}(\mu_r)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2) + E[\mu_r](\hat{sd}(\mu_j)^2)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2)) \\ &= \hat{\mu}_j(\hat{Var}(\mu_r)/(\hat{Var}(\mu_r) + \hat{sd}(\mu_j)^2)), \text{ since } E[\mu_r] \text{ is assumed to be 0.} \end{aligned} \tag{17}$$

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

[14] 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.

student is old for his grade, and whether the student is taking the course a year later than his peers at the school. It also includes post-graduation plans, parents' education categories, 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. We also allow for race matching effects between teacher and student, in acknowledgement of the findings of Hanushek et. al. (2005).

Observed prior inputs, $\mathbf{\tilde{Y}_i^{t-1}}$, 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 the impact of, for example, 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 2) was computationally intensive, and increased 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 run a fuzzy matching algorithm to match each teacher-class to a student-class. Since the grade level, race, and gender of each student in the student-class is observed, grade totals and race-gender cell totals can be constructed for the classes in the student-level data and compared to the corresponding grade totals and race-gender cell totals of the classes in the teacher-level data.[15] 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 4 describes the implementation of the fuzzy matching algorithm in detail.

The dataset began with 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, in the absence of a large number of students per transferring teacher, the number of transferring teachers connecting a school needs to be fairly large in order to reliably estimate the relative qualities of teachers at different schools. 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 each school is required to be connected to the network via a minimum of 3 transferring teachers (defined as a teacher who has taught at least 15 students at two different schools), 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 1 shows the distribution across schools of the number of connecting transfers. Figure 2 shows the distribution of exams administered across teachers in the sample. Figure 3 shows the number of students taught by each transferring teacher at the school at which he/she taught relatively few students. The latter figure illustrates that while many 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 and teachers who taught at unconnected schools, we are left with 4,016,343 test scores from 855,238 students and 18,498 teachers.

7 Results

7.1 Variance Decomposition

At first glance, there seem to exist considerable discrepancies across schools in test scores that differences in school quality, teacher quality, and teacher experience might be able to

[15] 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.

explain. Table 1 contains a decomposition of the variance in student test scores into within-school and between-school components. 8.7% of the variance is between schools, and 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). Furthermore, Table 2 shows that average teacher credentials also differ substantially across schools, creating the possibility that average teacher quality varies considerably 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%. However, a closer look at Table 1 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 and unobservable student and classroom characteristics, while row 7 shows that observable differences in student and classroom characteristics explains most of the between-school variance (75%). 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 inequitable 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 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, $d(\hat{exp})$, 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. These numbers, combined with the differences in average teacher experience across schools displayed in Row 4 of Table 2, 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 2 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 True Variance in Teacher Quality, School Sensitivity, and School Quality

Figure 4 displays the distribution of raw teacher effects $\{\hat{\mu}_r\}$. The distribution strikingly resembles a normal distribution, given that normality has not been imposed anywhere. While the standard deviation of $\hat{\mu}_r$ is .307, applying the measurement error correction described in section 5 leaves a true variance in μ_r of .030, with associated standard deviation of .174. An average student who receives a teacher who is at the 25th percentile of teacher quality can expect to move from the 50th percentile to the 45th percentile, while one who receives a teacher at the 5th percentile can expect to move down to the 38th percentile, assuming test scores are distributed normally.[16] This is substantial, and generally in line with most estimates found in the literature at the elementary and middle school level. Of course, some of this variation may reflect between-school differences in teacher quality that were not captured by previous studies. To investigate this possiblity, figure 6 plots the mean value of $\hat{\mu}_r$ at each school, weighting each teacher by the number of students he/she taught at that school. Applying the analogous measurement error correction (using the delta method to account for correlation in sampling error in $\hat{\mu}_r$ across teachers in the same school when calculating $sd(\overline{\mu_s})$) yields an estimate of the true between-school teacher quality variance of .0038, with associated standard deviation of .061. Since average teacher quality is nearly normally distributed as well, the estimates indicate that attending a school whose average teacher quality is in the 25th (5th) percentile moves an average student from the 50th percentile to the 48th (46th) percentile of the state test score distribution. So while average teacher quality does not vary dramatically across schools, attending a school with terrible teachers can still put a student at a meaningful disadvantage. Clearly, though, eliminating differences in average teacher quality across schools would not come close to eliminating test score gaps across schools.

Figures 8 and 9 display the distributions of raw and Bayesian posterior estimates of γ_s, respectively. 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, Figures 10 and 11 display the distributions of raw and Bayesian posterior estimates of δ, respectively. The estimated "true" standard deviation in δ is .112. If we do assume that the true mean teacher quality $E[\mu] = 0$ (see Appendix 2), and that there are no unobserved student inputs that cluster at the school level, then moving from a school at the 5th percentile of the δ distribution to the median school would increase

[16] This assumption is borne out by plots of the data.

expected test scores by .23 student-level standard deviations, 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. 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. Thus, any differences in the magnitude of effects of teacher quality 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). A likelihood ratio test rejects the hypothesis that $\gamma_s = 1 \ \forall \ s$ at the 95% level.

7.4 What Kinds of Students Attend Schools That Are Sensitive to Teacher Quality?

While having a sense of the distributions of average teacher quality, sensitivity to teacher quality, and school quality is useful in its own right, part of the motivation for estimating these distributions was to determine the extent to which students who are underprivileged in other dimensions are sorting into systematically different kinds of schools than better supported students. To this end, Table 4 provides the average values of $\hat{\overline{\mu}}_s$, $\hat{\gamma}_s$, 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 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 the most comprehensive measure of average student background, the average of an index, $X_{it}\beta + \tilde{Y}_i\alpha$, that weights student characteristics by how well they predict high school test score performance. 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 $\hat{\delta}_s$, displays 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, which replace $\hat{\delta}_s$ with $\hat{\gamma}_s$, offer more surprising results. Namely, 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 $(X_{it}B + \tilde{Y}_i\alpha)$ index of predicted test scores based on student observables have median sensitivity of 1.38, while those in the top quartile have median sensitivity of .70. This introduces the possibility of both equity and efficiency gains from the reallocation of high quality teachers from high performing to low performing schools.[17]

8 Counterfactuals

8.1 Efficient Allocation of Teachers to Schools

Given the nature of complementarity between school and teacher quality in the model, the efficient allocation of teachers to schools involves a positive assortative match in which the schools most sensitive to teacher quality are paired with the most effective teachers. We estimate the potential gain in state average test scores from the efficient allocation relative to the status quo, along with a scenario in which average teacher quality is equalized across schools, holding fixed the sorting of students to schools we observe in the data. To do this, 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.[18] Assuming multivariate normality, we can take 400 draws from

[17] To examine the sensitivity of the results to removing school-specific teacher development rates as a source of identifying variation for γ_s, we also estimated a version of the model in which schools are only differentially sensitive to the teacher-specific component of teacher quality, so that effective teaching quality is given by $\gamma_s\mu_r + d(exp_{rt})$. The results related to teacher quality change very little, but the variance in γ_s decreases by about X%. Low-performing schools are only about 15-20% more sensitive to teacher quality than the median school under this specification.

[18] We estimate the covariances in a manner analogous to the estimates of true variances in section 5. 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 ($\frac{1}{S}\sum_s cov(\epsilon_s^{\gamma}, \epsilon_s^{\delta})$).

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 (6).[19] Then, we construct the efficient allocation by sorting teachers by simulated quality and schools by simulated sensitivity to teacher quality, and 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 standard deviations, and reduces the standard deviation in test scores by 4.6%. Furthermore, the gap in average test scores between students in the top and bottom quartiles of the student background index $(X_{it}B + \tilde{Y}_i\alpha)$ is .1091 test score standard deviations smaller under the efficient allocation than under the status quo allocation. 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 across schools has almost no effect on the mean nor the variance of test scores relative to the status quo, 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 underperforming schools) would be likely to have the side effect of increasing average statewide test score performance.

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

Figures 5 and 7 display the distributions of the Bayesian posterior means of μ_r for each teacher and $\overline{\mu}_s$ for each school, respectively, assuming that μ_r and $\overline{\mu}_s$ are distributed nor-

[19] 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.

mally. These distributions allow us to think about the feasibility of teacher accountability systems that rely on student test score data. Suppose, for example, that North Carolina implemented a teacher accountability system that denied tenure to those in the bottom 5% of the posterior mean distribution of μ_r after four years of teaching. Interestingly, the projected efficacy of such a policy depends critically on which specification is used to calculate parameter estimates, and more importantly, standard errors. Using Bayesian posterior means and variances of teacher quality from the baseline specification, we find that 7.4% of those denied tenure under the above policy would in reality be above average teachers. Furthermore, 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 only .007 standard deviations, after adjusting for slightly larger class sizes (a tiny adjustment). A pertinent feature of this 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 considerable risk of unfairly removing effective teachers. However, if we construct the Bayesian posterior distributions 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 instead 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 only expect a .35 standard deviation increase in teacher quality, resulting in an overall average test score increase of .018 standard deviations. Recall from Table 3 that the estimated variance in teacher quality is nearly identical across the two specifications. This comparison shows that the reliability of value-added 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, given that 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 (even if they did not affect estimates of the true variance in teacher quality), the probability of correctly identifying an ineffective teacher may be even smaller than any model used so far would predict. Thus, 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.

8.3 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.[20] 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 much inequality 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, $\hat{\overline{\mu}}_i$.[21] 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:[22]

$$Var(\overline{\mu}_i) = Var(\hat{\overline{\mu}}_i) - (1/I)\sum_i (\sigma_i^{\mu})^2. \qquad (18)$$

Among students who took five tests, a one standard deviation increase in average teacher quality corresponds to an increase in average teacher quality of .062. In other words, a student whose average teacher is at the 5th (95th) percentile of the student-level average teacher quality distribution will have his test scores in each course reduced (increased) by an average of .141 standard deviations, solely by virtue of the teachers he was assigned

[20] 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.

[21] the results are similar if we condition on six or seven tests

[22] this method takes into account the correlation in sampling errors between $\hat{\mu}_r$ estimates associated with teachers from the same school

at his school.[23] 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 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, so we must make sure that none of this across-subject variation in teacher quality be included in a counterfactual estimate of the variance in student-level teacher quality under random assignment.[24] To overcome this problem, 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, to calculate the variance in student-level teacher quality under random assignment, 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.[25] Thus, while within-school variation in the average teacher quality experienced

[23] This calculation assumes a normal distribution, which is borne out by inspection.

[24] 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.

[25] 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 redid 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

across students does contribute to performance inequality, it does not seem to be the case that some kinds of students are systematically being assigned the inferior teachers at their schools; if anything, the variance in average teaching quality we observe is less than we would expect under random assignment!

Of course, verifying that large swaths of students are not systematically receiving inferior teachers does not rule out the possibility that a savvy student could learn considerably more if he/she knew which teachers at his/her school were relatively effective, and specifically chose time slots when these teachers were teaching. To examine this, I calculate the variance in within-student average teacher quality across feasible paths of teachers. Clearly, the payoff to selecting the best possible teachers depends critically on the size of one's school, since if there is only one teacher teaching in a given subject at a given time, there is no choice to be made! Thus, I sort student-specific estimates of the variance in average teacher quality across feasible paths, and find that a student whose variance places him at the 25th percentile can increase his expected average test score performance by only .017 standard deviations by choosing a path of teachers who is one standard deviation above the average across feasible paths for that student. In contrast, a student at the 75th (90th) percentile who chooses teachers who are one standard deviation above his/her feasible average teacher quality receives an expected payoff of .055 (.082) test score standard deviations. Thus, given uncertainty about which teachers are most effective, strategic course selection seems likely to be worth the trouble only for students at very large schools.

9 Testing the Exogenous Mobility Assumption

Consistent estimation of the parameters requires that teachers' location decisions are unrelated to ϵ_{it}. This is a strong assumption with important implications for the validity of the estimates. Fortunately, it is testable. Consider the following algorithm:

1. Select a teacher that transfers between two schools in the estimated network of schools.

2. Rather than imposing that this teacher has a single average quality μ_r, we treat him/her as if he/she were two distinct teachers that taught at two distinct schools

not sensitive to the method chosen, suggesting that either scheduling conflicts were rare, or most feasible paths were taken by some student.

with qualities μ_{r1} and μ_{r2}, and re-estimate the model. While we lose information about school quality and school sensitivity contained in the relative performance of the same teacher at two different schools, there remain nearly 3,000 other transfers in the network, so estimates of the other parameters should be affected only minimally.[26]

3. Repeat (1) and (2) for every transferring teacher.
4. Calculate the average before-after difference in teacher quality among transferring teachers: $\sum_{r\in\tilde{\mathcal{R}}} \hat{\mu}_{r2} - \hat{\mu}_{r1}$.

If the exogenous mobility assumption holds, the average within-teacher change in quality following a transfer should tend to zero as the number of transferrers gets large. I perform this test, and find that the average within-teacher change $(\hat{\mu}_{r2} - \hat{\mu}_{r1})$ is .0093, with an approximate standard error of .0079.[27] So we fail to reject the assumption of exogenous mobility at the 95% level. Indeed, only 51.4 percent of transferring teachers had higher quality estimates after a transfer than before $(\mu_{r2} > \mu_{r1})$. Given that the point estimate suggests the possibility of limited targeted mobility, one still must examine how such targeted mobility could occur, and how it might bias the estimates presented. Recall from (7) that ϵ_{it} can be decomposed into:

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

Recall also that Assumption 2 requires:

$$\begin{aligned}& E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred to } s' \text{ at some } t' < t] = 0 \\ & E[\epsilon_{ict}|s(i,t) = s', r(i,t) = r';\ r' \text{ transferred from } s' \text{ at some } t' > t] = 0 \qquad (19)\end{aligned}$$

Substituting the components in equation **??** for ϵ_{ict} in equation 19, we observe that a systematic relationship between a teacher's transfer decision and any of these components

[26] Of course, the downside to this method is that we are maintaining the exogenous mobility assumption with respect to the other transferring teachers in order to estimate the isolated teacher's effectiveness at each school.

[27] Calculating exact standard errors would require re-estimating the parameter variance matrix for each iteration of the above algorithm, which would require a prohibitive amount of computation. Thus, I assign to each $\hat{\mu}_{r1}$ and $\hat{\mu}_{r2}$ the mean standard error among teachers who teach a similar number of students as the transferrer taught at the school in question when calculating the standard error of the test statistic.

might produce the rejection we found. The first possibility 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.

A second possibility, 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.

A third possibility, related to κ_{rs}, is that teachers systematically transfer to schools at which they are relatively good 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. Note, however, that if all teachers can increase test scores by more at a given school, this is properly thought of as a school effect, and would be reflected in a higher δ for that school instead of being captured by κ_{rs}.[28]

A final possibility, 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.

One way to test for a violation related to this last possibility is to first re-estimate the model with school-year additive effects (so that $\tilde{\delta}_s$ is replaced with $\tilde{\delta}_{st}$).[29] Then, for each transferring teacher r, let $\tilde{t}(r)$ be the last year they teach at the school they transfer away from. We can compute the average value of $\tilde{\delta}_{st}$ for 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

[28] Similarly, if all above average teachers teach relatively well at a school, this suggests a high sensitivity to teacher quality (γ_s) at that school, and would not constitute a violation of the exogenous mobility assumption.

[29] 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.

these two measures across transferring teachers should be positive:

$$\sum_{r\in\tilde{\mathcal{R}}}(\sum_{t<=\tilde{t}(r)}\tilde{\delta}_{st}-\sum_{t>\tilde{t}(r)}\tilde{\delta}_{st})>0. \tag{20}$$

Likewise, for each transferring teacher, we can compute the average value of $\tilde{\delta}_{st}$ for the school he/she joined 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:

$$\sum_{r\in\tilde{\mathcal{R}}}(\sum_{t<=\tilde{t}(r)}\tilde{\delta}_{s't}-\sum_{t>\tilde{t}(r)}\tilde{\delta}_{s't})<0. \tag{21}$$

We perform this test, and find, strangely that while the schools that teachers join do indeed perform .009 student-level standard deviations better after the teachers join, and the schools teachers leave also perform .012 better after the transferrers leave. The conflicting outcomes of these tests indicate that movement based on changes in school quality is probably not driving the rejection of exogenous mobility in the test above.

Unfortunately, we are unable to distinguish between violations of the first three types. One important point, though, is that the impact of such endogenous mobility on the parameter estimates depends critically on the pattern of teacher mobility. Suppose that there is a job ladder, so that some undesirable schools tend to be net senders (generally replacing lost transferring teachers with novice teachers), while some desirable schools tend to be net receivers (losing teachers mainly to retirement, and replacing them with transferring teachers from other schools). If teachers tend to move to schools at which they have a comparative advantage, but only do so if it represents a step up the job ladder, then transferrers associated with undesirable schools will tend to be leaving teachers who were relatively bad fits for the school, while transferrers associated with desirable schools will tend to be joining teachers who are relatively good fits for the school. In this case, we will only observe transferrers at undesirable schools when they are teaching poorly relative to their own true talent, so that we will underestimate δ_s and overestimate the qualities of the other teachers at the school (and vice versa for the desirable schools). If the undesirable schools tended to be low quality and have low quality teachers, we would overestimate the variance in δ and underestimate the across-school variance in μ (but estimates of γ would be unbiased). On the other hand, suppose instead that such a job ladder did not exist, so that while teachers tended to move to comparative advantage (or move when they expect a change in their own qualities), each school sent and received approximately the same number of transferring teachers. In this case, half of the transferrers associated with each school would underperform relative to their own true talent, and the other half would overperform. This implies that the average values of κ_{rt}, ν_{sr}, and e_{ict} would tend to 0 among transferrers associated with each school as the number of transferrers got large. In this case, targeted mobility need not bias the estimates of any parameters.

Thus, one way to get a sense of the extent of bias in our estimates is to look for evidence of a job ladder in transfer patterns. 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.

Thus, we simulate a counterfactual distribution of the fraction of transferrers who are leavers at each school as follows. 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 (as 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 $\overline{f}^1, ..., \overline{f}^{329}$.

We also constructed a second counterfactual density in which a job ladder does exist. The method is the same, 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 12. 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 can reject 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. Thus, while targeted mobility may cause a bias in the $\overline{\mu}$ and δ estimates of some schools due to a small sample of transferrers who are disproportionately leavers or joiners, we have no reason to believe that a certain "type" of school is more or less likely to have biased estimates, or that our estimate of the between-school variance in teacher quality is biased. More generally, the absence of evidence for a job ladder suggests that teachers may not agree on a ranking of desirability for schools, or at least may not transfer based on any shared preferences that do exist.

10 Conclusion

In contrast to the horror stories recounted in the popular media in which the least privileged students attend disorganized schools with ineffective teachers, we find instead that quality teaching is fairly equitable 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.

The absence of 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 do not indicate the existence of a clear location ladder, suggesting instead the possibility that teachers hold weak or horizontal preferences among schools, so that the notion of universally "desirable" schools has been overstated.[30] Alternatively, given that teachers are hired by districts rather than schools, 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 ladders 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 magnify or mute the effectiveness or ineffectiveness of teachers. Interestingly, the schools that are most sensitive to teacher quality seem to be those with characteristics usually associated with low performance. Thus, policies that incentivize effective teachers to teach in struggling schools would be likely to increase average

[30] 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.

test scores statewide in addition to the clear equity payoff. Indeed, we estimate that the efficient allocation of teachers to schools would increase 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 indicate 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 suggest 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 critically 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. Thus, the pattern of teacher assignments observed in North Carolina does not indicate that students in more demanding curriculum tracks (or with particularly savvy and insistent parents) are systematically receiving the relatively effective teachers at their high schools. This finding can perhaps be partly attributed to the limited selection of teachers within a subject, particularly at small schools.

A couple of final 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 potentially less critical for across-school comparisons, it is nonetheless fairly easy to concoct stories in which it is violated. Secondly, the exogenous mobility assumption was tested and not rejected in my data. Thus, the extent of targeted mobility seems to be quite limited, and the pattern of mobility we observe (specifically, the absence of a job location ladder) 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 very cautious about generalizing these results to other states, grade levels, or outcomes.

11 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, 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

Boyd, D., Grossman, P., Lankford, H., Loeb, S., and Wyckoff, J. (2007). Who Leaves? Teacher attrition and student achievement. Working Paper 14022. Cambridge, MA: National Bureau of Economic Research.

Boyd, D., Lankford, H., Loeb, S., Wyckoff, J. (2005). *Journal of Policy Analysis and Management.* Vol. 24, No. 1, pp. 113-132.

Clotfelter, C. T., Ladd H. F., and Vigdor, J. L. (2007). “Teacher Credentials and Student Achievement: Longitudinal Analsis with Student Fixed Effects.” *Economics of Education Review.* Vol. 26, No. 6. pp. 673-682.

Goldhaber, D., B. Gross and D. 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, E.A., Rivken, S.G., and Kain, J.F. (2005). “Teachers, Schools, and Academic Achievement.” *Econometrica.* Vol. 73, No. 2.

Hanushek, E., Kain, J. F., O’Brien, D., and Rivkin, S. G. (2005). “The Market for Teacher Quality” Working Paper 11154. Cambridge, MA: National Bureau of Economic Research.

Kane, T., J. Rockoff and D. 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.

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.

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.

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.

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.

12 Tables

Table 1: 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 $Var(X_{ict}\beta_c + Y_i^{t-1})$	.514	.717	.569
(3)	Effective School and Teacher Quality $Var(\lambda_{srj})^*$	.047	.216	.052
(4)	Cov(Stu. Background, Eff. Sch./Tch. Qual.) $2 * Cov(X_{ict}\beta_c + Y_i^{t-1}, \lambda_{srj})$	.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 $Var(\overline{X_s\beta_c} + \overline{Y}_s^{t-1})$	.058	.242	.065
(8)	Total School Quality $Var(\overline{\lambda}_s)^{**}$	.009	.097	.010
(9)	Cov(Avg. Stu. Background, Total Sch. Qual.) $2 * Cov(\overline{X_s\beta_c} + \overline{Y}_s^{t-1}, \overline{\lambda}_s)$	.010	–	.011

*λ_{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 3). Note that $\lambda_{srj} = \lambda_h = \delta_s + \gamma_s(\mu_r + d(ex_{rt})) + \omega_{srj}$. Thus, $Var(\lambda_{srj})$ represents the variance between school-teacher-experience cells, which 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.

Table 2: 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 3: Raw and Error-Adjusted Variances in μ_r, $\overline{\mu}_s$, γ_s, and δ_s: Baseline and Linear Specifications

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

	Mean Student Characteristic		Mean Teacher Quality ($\hat{\bar{\mu}}_s$)		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 ($X_i\hat{\beta} + Y_i^{t-1}\hat{\alpha}$)	-.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 ($\hat{\bar{\mu}}_s$) among schools in either the top or bottom quartile of schools sorted by their values of the student background measure associated with a given row.

Mean 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.

13 Figures

Figure 1: Distribution of the Number of Transferrers Across Schools

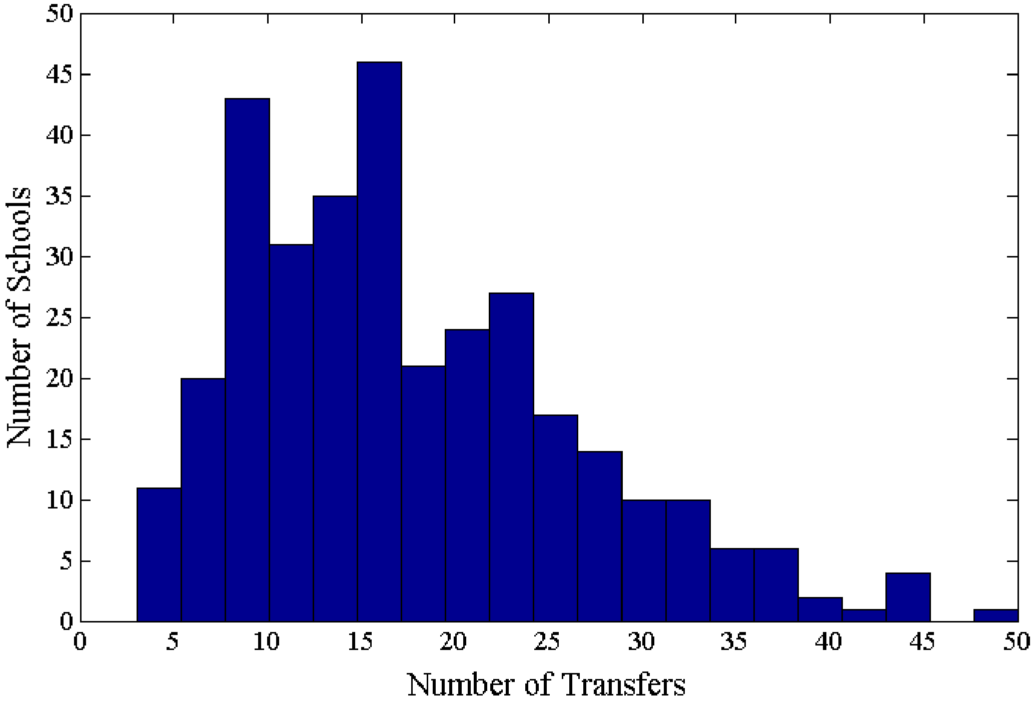


Figure 2: Distribution of Number of Exams Administered Across Teachers

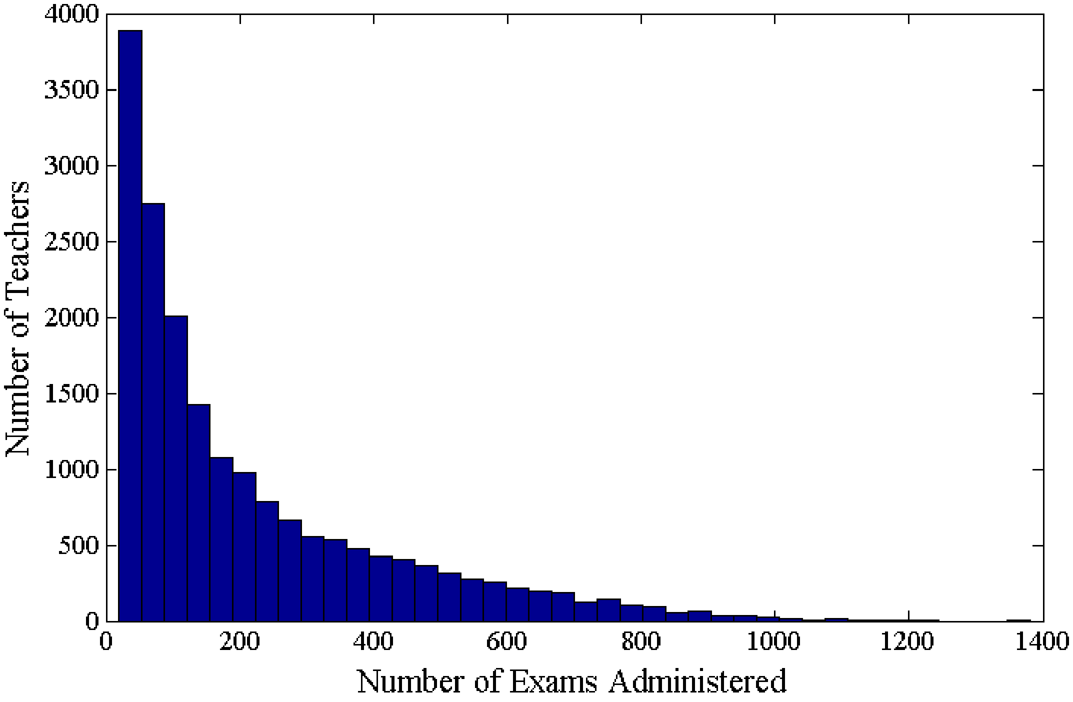


Figure 3: Distribution of Min(Total Students$_1$, Total Students$_2$) for Transferring Teachers

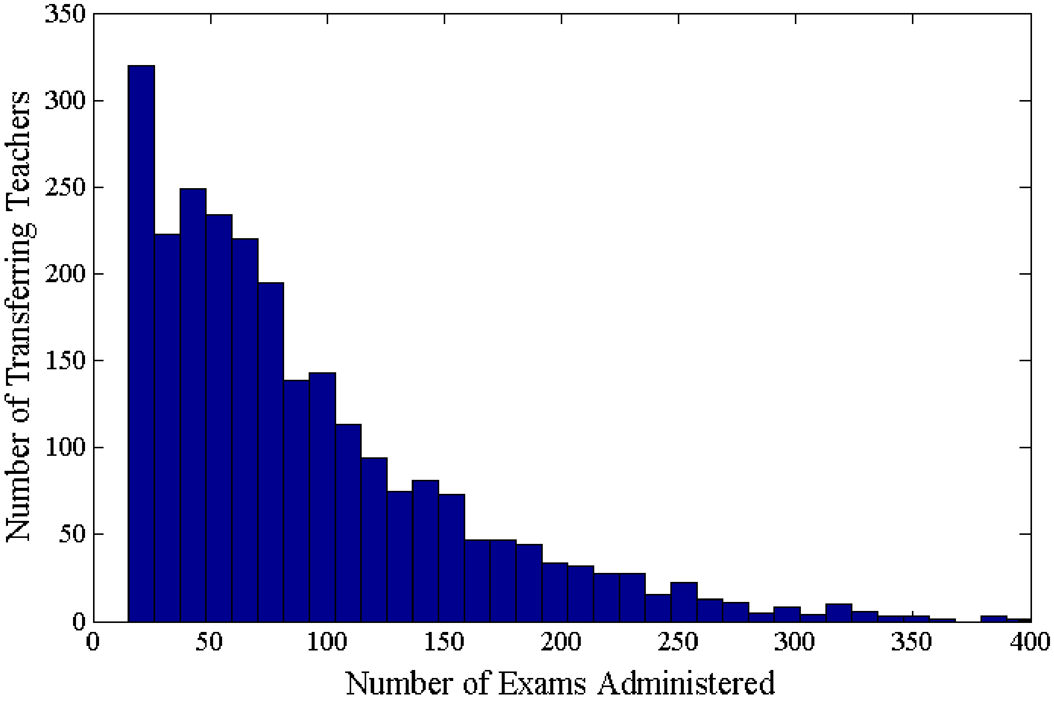


Figure 4: The Distribution of Unadjusted Teacher Quality Estimates ($\hat{\mu}$)

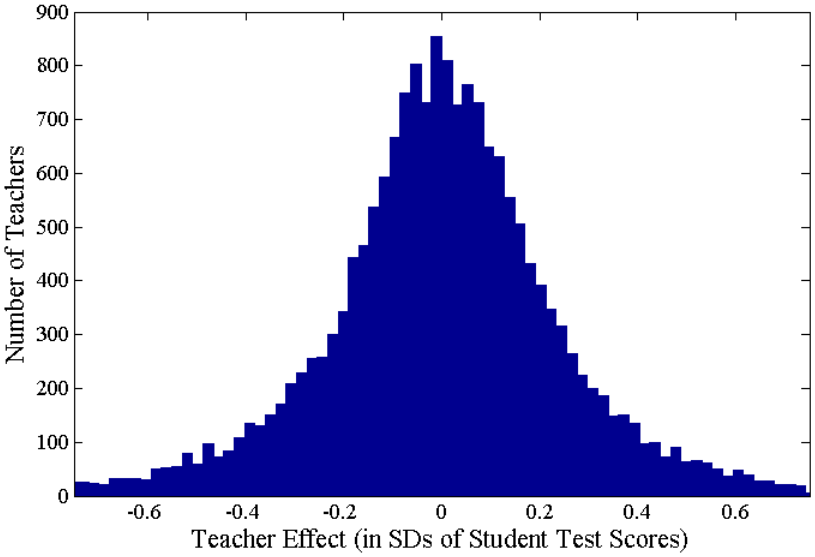


Figure 5: The Distribution of Bayesian Posterior Means of Teacher Quality (μ^B)

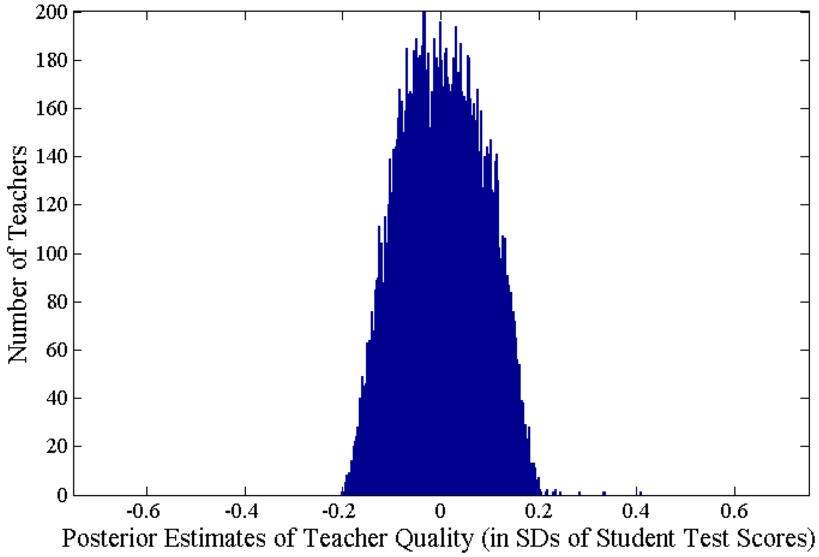


Figure 6: The Distribution of Unadjusted School Average Teacher Quality Estimates ($\hat{\mu}$)

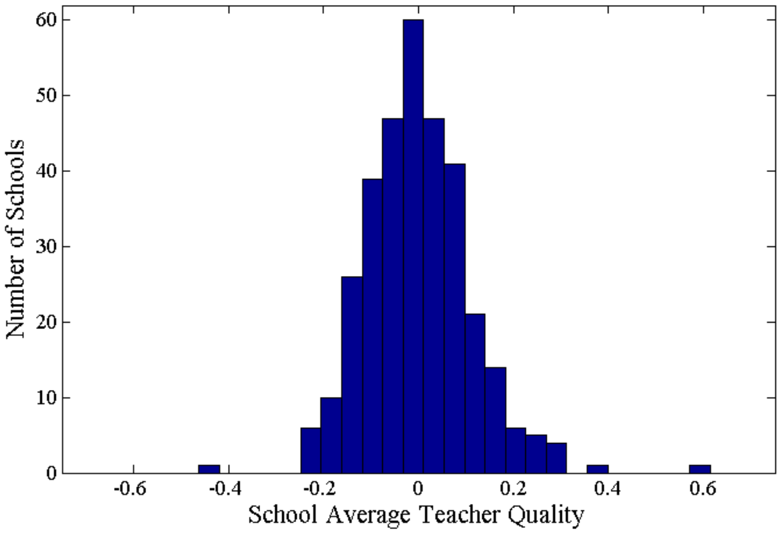


Figure 7: The Distribution of Bayesian Posterior Means of Average Teacher Quality Across Schools ($\overline{\mu}^B$)

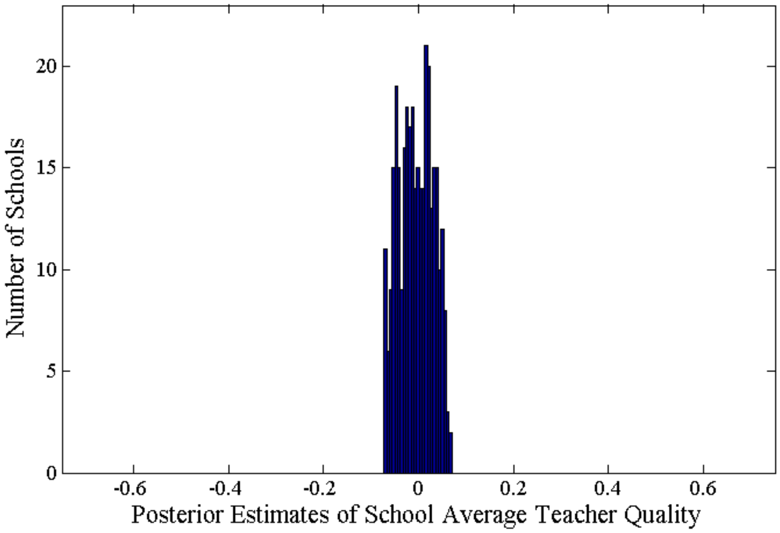


Figure 8: The Distribution of Unadjusted School Sensitivity Estimates ($\hat{\gamma}$)

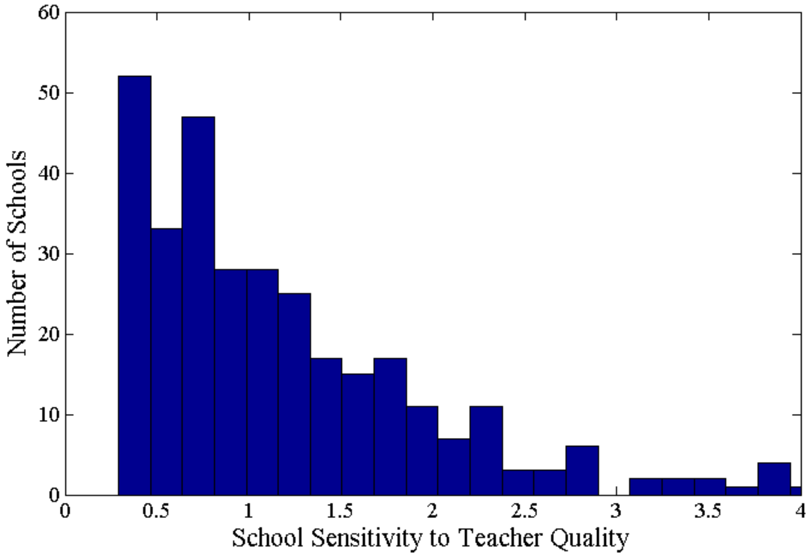


Figure 9: The Distribution of Bayesian Posterior Means of School Sensitivity (γ^B)

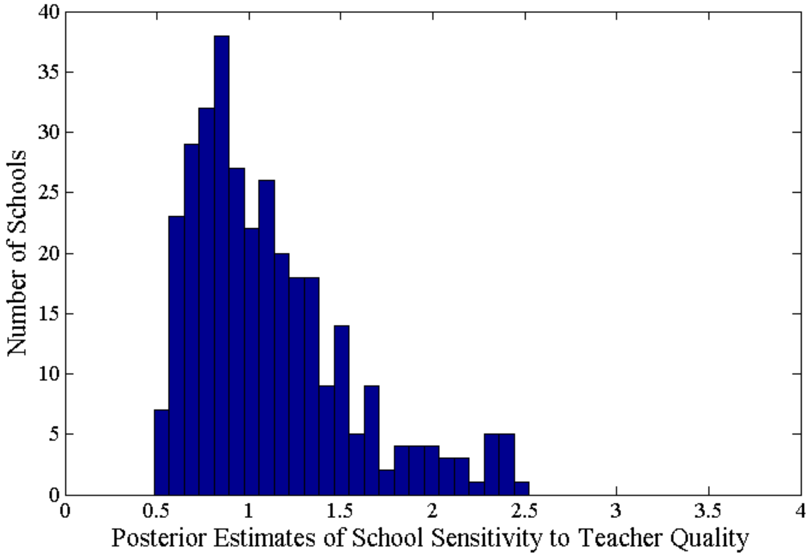


Figure 10: The Distribution of Unadjusted School Quality Estimates ($\hat{\delta}$)

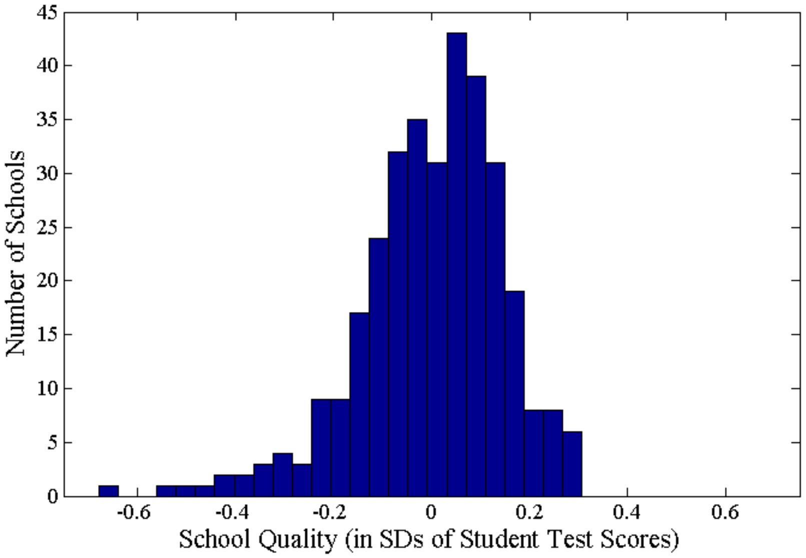


Figure 11: The Distribution of Bayesian Posterior Means of School Quality (δ^B)

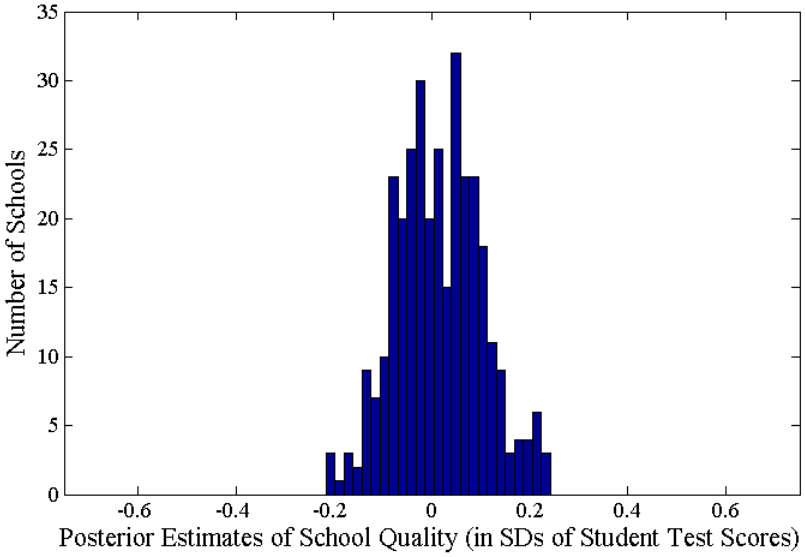
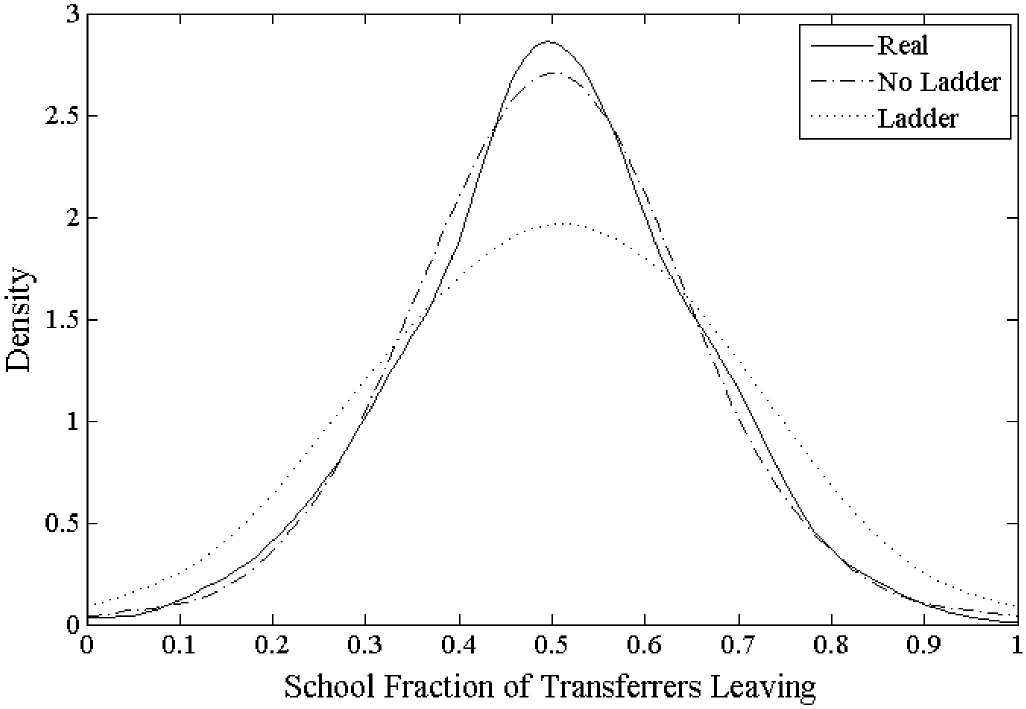


Figure 12: Testing for the Existence of a Job Ladder: A Plot of the Distribution Across Schools of the Fraction of Associated Transferring Teachers That Are Leavers (vs. Arrivers) Using Sample Data, Simulation with No Ladder, and Simulation with Ladder



Appendix

1 Proof of Identification

Proposition:

Consider a set $\mathcal{S}$ of S schools and a set $\mathcal{R}$ of R teachers, 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'. Finally, suppose that $\gamma_s \neq 0$ for any $s \in \mathcal{S}$. 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:

Consider two teachers, R_{11} and R_{21}, who teach in schools S_1 and S_2, respectively. R_{11} is observed at the experience levels x and x' in S_1, and R_{21} is observed at the same experience levels x and x' in school S_2. Let $Z_{ict} = Y_{ict} - X_{ict}\beta_c - \tilde{Y}_i^{t-1}\alpha_c$.[31] Suppose without loss of generality that each teacher is observed at experience levels x and x' at times t and t', respectively. Comparing the performance of students taught by R_{11} at time t with students taught by R_{11} at time t', we have:

$$E[Z_{ict}|s(i,t) = S_1, r(i,c,t) = R_{11}] - E[Z_{ict'}|s(i,t') = S_1, r(i,c,t') = R_{11}] = \gamma_1[d(x) - d(x')] \quad (22)$$

Making the same comparison for teacher R_{21}, we have:

$$E[Z_{ict}|s(i,t) = S_2, r(i,c,t) = R_{21}] - E[Z_{ict'}|s(i,t') = S_2, r(i,c,t') = R_{21}] = \gamma_2[d(x) - d(x')] \quad (23)$$

[31] note that β and α can be identified separately from within teacher-year variation.

Taking the ratio of these two moments gives:

$$\frac{E[Z_{ict}|s(i,t)=S_2,r(i,c,t)=R_{21}]-E[Z_{ict'}|s(i,t')=S_2,r(i,c,t')=R_{21}]}{E[Z_{ict}|s(i,t)=S_1,r(i,c,t)=R_{11}]-E[Z_{ict'}|s(i,t')=S_1,r(i,c,t')=R_{11}]}$$
$$=\frac{\gamma_2[d(x)-d(x')]}{\gamma_1[d(x)-d(x')]}=\frac{\gamma_2}{\gamma_1} \tag{24}$$

Thus, if we normalize $\gamma_1 = 1$, we can identify the relative sensitivity of S_2, γ_2, since $\gamma_1 \neq 0$ and $d(x) - d(x') \neq 0$ by assumption. 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.

Reconsider equation 22 above. With the distribution of $\{\gamma^s\}$ 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 $d(x)$ by comparing teachers' performance in subsequent years to their performance in the first year.[32]

Now, consider the level of student performance of teacher R_{11} with experience x in an arbitrarily chosen year, t:

$$E[Z_{ict}|s(i,t)=S_1,r(i,c,t)=R_{11}]=\delta_1+\gamma_1(d(x)+\mu_{11}) \tag{25}$$

If we normalize $\delta_1 = 0$, the level of performance of R_{11} while at S_1 identifies μ_{11}. Now, let R_{12} be a second teacher who taught at S_1 at time t, in this case at experience level x''. Comparing the performance at time t of R_{12} with R_{11}, we have:

$$E[Z_{ict}|s(i,t)=S_1,r(i,c,t)=R_{12}]-E[Z_{ict}|s(i,t)=S_1,r(i,c,t)=R_{11}]$$
$$=\gamma_1(d(x'')-d(x)+\mu_{12}-\mu_{11}) \tag{26}$$

Since d(*), μ_{11}, and $\{\gamma^s\}$ have already been identified, this difference moment identifies μ_{12}. Similar comparisons identify the qualities of all teachers at school S_1, $\{\mu_{1*}\}$.

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. Comparing her performance

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

while at S_k with another teacher at the school, R_{k1}, in an arbitrary year t' with experiences x'' and x''' respectively, we have:

$$\begin{aligned} &E[Z_{ict'}|s(i,t') = S_k, r(i,c,t') = R_{1j}] - E[Z_{ict'}|s(i,t') = S_k, r(i,c,t') = R_{k1}] \\ &= \gamma_k(d(x'') - d(x''') + \mu_{1j} - \mu_{k1}) \qquad (27) \end{aligned}$$

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 level of test scores for any teacher l in a given year at school k gives:

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

which identifies δ_k. 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 and $S-1$ 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. Suppose (w.l.o.g) that for estimation γ_l^e is normalized to 1, δ_l^e is normalized to 0, and $d^e(0)$ is normalized to 0, where the "e" superscript indicates the estimated value before final normalization.[33] 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:

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

[33] Normalizing so that the parameters average to 0 before estimation eliminates the sparsity of the design matrices.

Consider a second teacher, h, with the same experience, and imagine 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 have, using the above approximation:

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

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

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:[34]

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

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

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

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

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

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:

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

[34] 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 schools have sensitivity greater than 1.

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:

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

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:

$$\delta_l + \gamma_l[\mu_j + d(0)] \approx \delta^e_l + \gamma^e_l[\mu^e_j + d^e(0)] = \mu^e_j \tag{38}$$

Revisiting equation 31, and substituting for μ^e_j and γ^e_m, we have:

$$\gamma^e_m(\mu^e_j - \mu^e_h) = \frac{\gamma_m}{\gamma_l}(\delta_l + \gamma_l[\mu_j + d(0)] - \mu^e_h) \approx \gamma_m(\mu_j - \mu_h) \tag{39}$$

Solving for μ^e_h gives:

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

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:

$$\mu^e_h - \mu^e_j = \gamma_l(\mu_h - \mu_j) \tag{41}$$

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

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

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:

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

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 nonlinear 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 the true average teacher with no experience neither increases nor decreases test scores: $\sum_r(\mu_r)+d(0)=0$. Then, differences in estimated school additive effects can be interpreted as differences in the two schools' abilities to increase test scores. The equation above reveals, however, that a different arbitrary assumption regarding decomposition of the level of state average test scores would result in a different variance in school additive effects. Given that δ is already an amalgam of actual school effects and average unobservable student quality, from this point forward we will focus very little on the estimates of δ.

3 Details of Estimation

Recall that we are estimating the following equation:

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

where:

Y is a vector of N test scores, aggregated across classes, courses, schools, and years.

X is an NxK matrix of covariates. Some covariates are at the classroom level and 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.

$\tilde{\mathbf{Y}}$ is an NxL matrix of prior test scores and squares of prior test scores.

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.

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.

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 vector of J parameters that indicates how much an average teacher in the corresponding experience cell increases test scores, relative to a first year teacher.

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 6 can be rewritten in the following way:

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

where

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

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:

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

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:

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

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 $e\tilde{x}p(i,j) = 1$ 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:

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

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.[35]

4 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 over come 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

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

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 4 grade levels, and number in each of 6 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.[36]

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.

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.[37]

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. [38]

[36] Note that this implies that we don't 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.

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

[38] 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

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 for the match to be verified, the fraction of classrooms matched ranged from 50% in Algebra 1 to 85% in Physics.[39]

5 Teachers and Schools: Complements or Substitutes?

The model we have estimated permits investigation into whether schools and teachers are primarily complements or substitutes. If they are primarily complements, then improving schools requires both an emphasis on improving teaching quality as well as an emphasis on providing a school environment in which good teachers can be effective. If instead schools and teachers are primarily substitutes, then we can make progress toward closing

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.

[39] 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.

test score gaps simply by either improving poor school environments with a given set of teachers or by encouraging excellent teachers to transfer to underperforming schools, without intervening in the schools themselves. Thus, in this section, we attempt to evaluate the extent of complementarity between teachers and schools by isolating the two components of the student test score variance that are attributable directly to schools: a "substitute" component due to δ_s, and a "complement" component due to γ_s. Since δ_s is an additive component, the substitute component of test score variance is straightforward: $Var(\delta_s)$, which we calculated above as .013. However, the effect of the γ_s parameters on student test scores depends not just on $Var(\gamma_s)$, but also on the variance in teacher quality, $Var(\mu_r)$, its between-school component, $Var(\overline{\mu}_s)$, and the extent to which high sensitivity schools have high average teacher quality.[40] To sidestep these issues, we focus on the hypothetical case in which teachers are randomly assigned to schools, so that average teaching quality at each school is approximately the same ($Var(\overline{\mu}_s) = 0$), and γ_s and $\overline{\mu}_s$ are independent. In this context, the variance in effective teacher quality, $Var(\gamma_s\mu_r)$, is given by:

$$Var(\gamma_s\mu_r) = E[\gamma_s]^2 Var(\mu_r) + E[\mu_r]^2 Var(\gamma_s) + Var(\mu_r)Var(\gamma_s) \tag{52}$$

First, notice that due to our normalization of the scale of teacher quality, $E[\mu_r]^2 = 0$. Second, note that even if all schools were equally sensitive to quality ($Var(\gamma_s) = 0$), the variance in effective teacher quality would still be $E[\gamma_s]^2 Var(\mu_r) = Var(\mu_r)$. Thus, the "complement" component can be defined as:

$$E[\gamma_s]^2 Var(\mu_r) + Var(\mu_r)Var(\gamma_s) - Var(\mu_r) = Var(\mu_r)(E[\gamma_s]^2 + Var(\gamma_s) - 1) \tag{53}$$

Then, given the estimated variance in teacher quality $Var(\mu_r)$, if teachers were randomly assigned to schools then the fraction of school-attributable variance that comes from the complementary component would be given by:

$$\frac{Var(\mu_r)(E[\gamma_s]^2 + Var(\gamma_s) - 1)}{Var(\delta_s) + Var(\mu_r)(E[\gamma_s]^2 + Var(\gamma_s) - 1)} \tag{54}$$

[40] This is partly captured by $Cov(\overline{\mu}_s, \gamma_s)$, but higher moments of the joint distribution of $\overline{\mu}_s$ and γ_s also contribute to the variance in effective teaching quality, $Var(\gamma_s\mu_r)$.

Note that $Var(\mu_r) = .030$, $Var(\gamma_s) = .320$, $Var(\delta_s) = .013$ and given that the $\{\gamma_s\}$ are nearly log-normally distributed, $E[\gamma_s]^2 = 1.11$. Evaluating the formula above using these parameters, we find that almost exactly half (.505) of the school-attributable test score variance is due to the complementary component. Thus, while school and teaching inputs do partly substitute for each other, they seem to exhibit substantial complementarity as well.

6 Calculation of Standard Errors

Mimicking the estimation procedure documented in Appendix 3, 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 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 $(X'X)$ 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:

$$A = A^1 + A^2 + ... + A^N \tag{55}$$

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:

$$AG = A^1G + A^2G + ... + A^NG \tag{56}$$

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.