2014

Teacher Quality and Student Inequality - Web Appendix (Revised 2014)

Richard K. Mansfield
Cornell University, rm743@cornell.edu

Teacher Quality and Student Inequality - Web Appendix (Revised 2014)

Abstract
This appendix is a supplement to the author's paper, Teacher Quality and Student Inequality, which can be found here: http://digitalcommons.ilr.cornell.edu/workingpapers/162/

Keywords
education, teachers, student, test score performance, public high schools

Disciplines
Junior High, Intermediate, Middle School Education and Teaching

Comments
Suggested Citation
Mansfield, R. K. (2014). *Teacher quality and student inequality: Web appendix* [Electronic version]. Retrieved [insert date], from Cornell University, School of Industrial and Labor Relations site: http://digitalcommons.ilr.cornell.edu/workingpapers/174/

TEACHER QUALITY AND STUDENT INEQUALITY

Web Appendix

Richard K. Mansfield

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

A1 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 grouped by subject and 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/level combination in a year, so any biology classroom in the EOC dataset can be safely attributed to the single biology teacher. In large schools, there may be four biology teachers, each teaching four sections, making this process much more subtle.

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

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

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

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

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

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

6. If more than one classroom in the SAR dataset is matched to a given classroom in the EOC dataset at a given standard, but the teacher is the same in each of the SAR classrooms, that teacher is matched to the EOC classroom.[1]

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

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

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.

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. [2]

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. I 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 I wanted an algorithm that would not punish too severely 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 83% for Algebra 1 to 95% for Chemistry (since fewer people take Chemistry, there are many fewer sections and teachers, making it much easier to match). If I 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

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

be verified, the fraction of classrooms matched ranged from 50% in Algebra 1 to 85% in Physics.[3] Web Appendix Table A1 displays match rates by subject, while Web Appendix Table A2 displays match rates by quartiles of school size and average student background.

A2 TEST-SCORE SCALING

The test scores are scaled scores that I re-standardized to have zero mean and unit variance within each subject-year combination. Meghir and Rivkin (2010) have noted that if monotonic transformations of test scores convey the same information about student learning, a particular form of the education production function may be specific to an arbitrarily chosen test score scale. This is of potential concern in this context, since the method relies upon pooling tests from different subjects and years. However, most widely-used standardized tests have undergone considerable field-testing and analysis using item response theory to ensure that the final test instrument properly evaluates mastery of the relevant content and effectively differentiates between students throughout the distribution of learning. Consequently, assuming that all such tests, once standardized, share a common relationship with teacher and school inputs is not unreasonable. Thus, one way to interpret the estimates is that they measure the impacts of teachers or schools on standardized z-scores from field-tested exams designed for the whole range of student abilities, and would be general to any test fitting this description. Applying convex and concave monotonic transformations of test scores and re-estimating has little impact on the results below, except in cases where such transformations are so extreme that they introduce floor or ceiling effects. Web Appendix Figures A1-A3, which plot histograms for the tests associated with each subject-year, show that the tests used in this analysis are not plagued by floor effects nor ceiling effects.[4] One reason why the results are less sensitive to alternative scalings is that there is considerable overlap in the academic background of students taught by different teachers. Only 24% of the variance in the regression index of student

3. 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.

4. See Koedel and Betts (2010) for an analysis of the impact of floor and ceiling effects on value-added estimates.

background $X_{it}\beta + \tilde{Y}_i\alpha$ consists of variation in average background of students between different teachers. Thus, scalings that emphasize gains in different parts of the achievement distribution produce very little reshuffling of teacher rankings, since most teachers have a mix of well- and poorly-prepared students. In the extreme case in which all teachers taught a sample of students representative of the full statewide distribution of academic preparation and ability, the ranking of teachers would be completely insensitive to alternative scalings (at least in the absence of teacher comparative advantages for particular student ability types).

A3 CALCULATION OF STANDARD ERRORS

Standard errors are calculated using the standard sandwich formula, $V = (G'G)^{-1}G'\Omega G(G'G)^{-1}$, where in our context G is the horizontal concatenation of the vectors of student charateristics and prior test scores along with design matrices for school-course fixed effects, teacher fixed effects, and experience cell fixed effects, and $\Omega = var(\epsilon)$. However, the relative simplicity of the variance formula belies the considerable computational difficulty associated with its evaluation. Recall that there are $N = 4,016,964$ test-score level observations, 840 subject-specific coefficients on student background characteristics and prior test scores, $3,357$ school-subject fixed effects, and $19,826$ teacher fixed effects and prior test scores, and 5 experience cell effects, resulting in $H = 24,030$ parameters. Thus, direct evaluation of V would require both the inversion of a $24,030$ x $24,030$ matrix ($G'G$) and the construction of a 4 million x 4 million matrix (Ω). Both of these operations exceed the memory limits of even very powerful servers. A couple of subtle tricks were necessary to make this calculation feasible within a reasonable length of time. First, note that V can be written as the product $V = AGB$, where A is the H x N matrix $(G'G)^{-1}G'\Omega$, G is N x H, and B is the H x H matrix $(G'G)^{-1}$. Next, let $A(k)$ denote the k-th column of A, and define A^k as the H x N matrix in which the k-th column consists of $A(k)$, and all other elements are zeros. Note that A

can be written as:

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

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:

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

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.

A4 MEASUREMENT ERROR

A4.1 Error-Corrected Variance in Teacher Quality

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_{ict}. This section describes the procedure used to isolate the true variance in teacher quality.

First, I define each estimated teacher fixed effect $\hat{\mu}_r$ as the sum of the teacher's true quality and an uncorrelated error component: $\hat{\mu}_r = \mu_r + \xi_r$. Let $\mathcal{ICT}$ be the set of student-course-year combinations, and let N be the total number of such combinations (i.e. the number of observations in the dataset). Then the sample student-weighted variance in estimated teacher quality can be decomposed as:[5]

$$\frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\hat{\mu}_{r(i,c,t)})^2 = \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\mu_{r(i,c,t)})^2 + \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\xi_{r(i,c,t)})^2 \tag{3}$$

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

$$\hat{Var}(\mu_r) = \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\hat{\mu}_{r(i,c,t)})^2 - \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\xi_{r(i,c,t)})^2. \tag{4}$$

ξ_r is not observed, but

$$\frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\xi_{r(i,c,t)})^2 \approx \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}E[(\xi_{r(i,c,t)})^2] = \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(se(\xi_{r(i,c,t)}))^2, \tag{5}$$

5. I treat the set of teachers in North Carolina in the 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 I calculate can thus be interpreted as the variance in the true quality of teachers teaching in the data, which covers all teachers in nearly all high schools in North Carolina over an eleven year period.

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

$$\hat{Var}(\mu_r) = \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\hat{\mu}_{r(i,c,t)})^2 - \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}}(\hat{se}(\xi_{r(i,c,t)}))^2 \tag{6}$$

I use the same technique to estimate the true variance in school average teacher quality $\overline{\mu}_s$.

A4.2 Shrinkage Estimates of School-Course-Year Averages of Teacher Quality

Note that the true variance in school-course-year means of teacher quality, $Var(\overline{\mu}_{sct})$ is different from $Var(\mu)$ for two reasons. First, to the extent that teachers of similar quality sort into the same schools (and perhaps into the same courses or at the same points in time) the true test-score-weighted variance among teachers observed teaching in the same school-course-year combination is less than the overall variance. Second, averaging among multiple teachers drawn from the within school-course-year distribution of teacher quality further reduces $Var(\overline{\mu}_{sct})$.

I calculate $Var(\overline{\mu}_{sct})$ using an analogous measurement error adjustment procedure to the one described in the previous subsection. In order to use the same technique, I need to first calculate the true variance in teacher quality within school-course-year combinations. Denote a teacher's deviation in quality from the student-weighted mean among those teaching in the same school-course-year by $\tilde{\mu}_r = \mu_r - \sum_{(i,c,t)\in\mathcal{ICT}(r)}\mu_{r(i,c,t)}$, where $\mathcal{ICT}(r)$ is the set of students taught in the same school-course-year combination as teacher r. Following the same logic as presented in the previous subsection, I can estimate the true within school-course-year variance in teacher quality via:

$$\hat{Var}(\tilde{\mu}_r) = \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}(r)}(\hat{\tilde{\mu}}_{r(i,c,t)})^2 - \frac{1}{N}\sum_{(i,c,t)\in\mathcal{ICT}(r)}(\hat{se}(\tilde{\xi}_{r(i,c,t)}))^2 \tag{7}$$

Note that implementing this formula requires the use of appropriate standard errors for estimates of teacher-specific deviations in quality relative to the school-course-year mean, $\hat{se}(\tilde{\xi}_r)$. Since the

sampling error in estimates of $\hat{\mu}_r$ among teachers from the same school, course, and year will be correlated, I use the delta method combined with the full covariance matrix of original estimates to calculate standard errors for $\hat{\tilde{\mu}}_r$. I then treat each school-course-year mean teacher quality as if it were a draw from a distribution of school-course-year means that were each constructed using the same number of teachers as the mean in question, along with the same fraction of students taught by each teacher. Thus, the true variance of the distribution from which a particular school-course-year mean is drawn will be specific to the number of teachers contributing to the mean and their shares of total students taught. Specifically,

$$(8) \qquad Var(\overline{\mu}_{sct})_{(sct)'} = Var(\sum_{r \in (\mathcal{SCT})'} w_r \mu_r) = \sum_{r \in (\mathcal{SCT})'} Var(w_r \mu_r) = Var(\tilde{\mu}_r) \sum_{r \in (\mathcal{SCT})'} w_r^2$$

where w_r is the share of students taught in school-course-year combination sct that were taught by teacher r.

This procedure imposes that the true deviations in quality relative to the overall school-course-year mean of different teachers are independent. Once estimates of the signal strength $Var(\overline{\mu}_{sct})_{(sct)'}$ has been calculated for each school-course-year mean, I calculate the variance in the sampling error for each school-course-year mean, $Var(\nu^{\mu}_{sct})$ using the square of the standard error for $\hat{\overline{\mu}}_{sct}$. This standard error also requires use of the delta method to account for correlation in sampling error among estimates of teacher quality for teachers teaching in the same school, course, and year. Combining the signal and noise components yields empirical Bayes posterior mean estimates of school-course-year mean teacher quality:

$$(9) \qquad \mu^{EB}_{(sct)'} = \frac{\hat{Var}(\overline{\mu}_{sct})_{(sct)'}}{\hat{Var}(\overline{\mu}_{sct})_{(sct)'} + \hat{Var}(\nu^{\mu}_{(sct)'})} \hat{\overline{\mu}}_{(sct)'}$$

A5 TESTING FOR VIOLATIONS OF THE EXOGENOUS MOBILITY ASSUMPTION

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

Specifically, recall that Assumption 2 requires:

$$
\begin{aligned}
&E[\epsilon_{ict}|r(i,c,t) \in \tilde{\mathcal{R}}_{(s',c')}, (s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \\
&= E[\epsilon_{ict}|(s(i,t),c) = (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; (s',c') \in \mathcal{SC} \\
&E[\epsilon_{ict}|r(i,c,t) \in \tilde{\mathcal{R}}_{(s',c')}, (s(i,t),c) \in \mathcal{SC} \setminus (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \\
&= E[\epsilon_{ict}|(s(i,t),c) \in \mathcal{SC} \setminus (s',c'), \tilde{\mathbf{Y}}_i^{t-1}, \mathbf{X}_{ict}] \; \forall \; (s',c') \in \mathcal{SC}, \qquad (10)
\end{aligned}
$$

Substituting the specification of the error components in for ϵ_{ict} in (10), we observe that a systematic relationship between a teacher's transfer decision and any of these components would violate Assumption 2. However, I restrict attention to two mechanisms that are particularly plausible, and develop tests for each.[6]

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

A5.1 Do Teachers Try to Escape from Declining Schools?

The first mechanism, related to ϕ_{st}, is that teachers systematically transfer toward or away from schools that are about to get better or worse, relative to the school's average quality over the sample period. This might occur, for example, if teachers follow a particularly effective principal when he or she moves from school to school.

To test this hypothesis, I re-estimate the model with school-year additive effects, so that δ_{sc} is replaced with δ_{st}.[7] Then, for each transferring teacher r, let $\tilde{t}(r)$ be the last year they teach at the school they transfer away from (denoted s). We can compute the average value of $\hat{\delta}_{st}$ for the school he/she left for the years during/before their exit ($t <= \tilde{t}(r)$) and for the years after they left ($t > \tilde{t}(r)$). If teachers are transferring away from schools that are about to decline, then the mean difference among these two measures across transferring teachers should be positive:[8]

$$\frac{1}{|\tilde{\mathcal{R}}|}\sum_{r\in\tilde{\mathcal{R}}}\left(\frac{1}{|\tilde{\mathcal{T}}_r^B|}\sum_{t<=\tilde{t}(r)}\hat{\delta}_{st}-\frac{1}{|\tilde{\mathcal{T}}_r^A|}\sum_{t>\tilde{t}(r)}\hat{\delta}_{st}\right)>0. \tag{11}$$

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

$$\frac{1}{|\tilde{\mathcal{R}}|}\sum_{r\in\tilde{\mathcal{R}}}\left(\frac{1}{|\tilde{\mathcal{T}}_r^B|}\sum_{t<=\tilde{t}(r)}\hat{\delta}_{s't}-\frac{1}{|\tilde{\mathcal{T}}_r^A|}\sum_{t>\tilde{t}(r)}\hat{\delta}_{s't}\right)<0. \tag{12}$$

I perform these tests and find, strangely, that while the schools that teachers join do indeed perform .009 student-level standard deviations better after the teachers arrive, the schools teachers

7. 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 I already have verified the existence of a connected graph between schools.

8. $|\tilde{\mathcal{R}}|$ denotes the size of the set of transferring teachers, $|\tilde{\mathcal{T}}_r^B|$ denotes the total number of years in the sample spent by transferring teacher r at school s before transferring, and $|\tilde{\mathcal{T}}_r^A|$ denotes the total number of years in the sample at school s after r transferred away.

leave also perform .012 better after the transferrers leave. The small and statistically insignificant magnitudes and conflicting interpretations of these test statistics suggest that teacher mobility is generally not driven by changes in school quality.

A5.2 Do Teachers Move to Schools Where They Are Better Matched?

The education production function employed to this point has assumed away the existence of an idiosyncratic match quality between particular schools and teachers. However, there is a growing literature inspired by Abowd, Kramarz, and Margolis' (1999) decomposition of wages that models employer-employee sorting based on match quality and its implications for the interpretation of firm and worker fixed effects (see Lise, Meghir, and Robin [2008] and Lopes de Melo [2009] for examples). Some of the issues raised in this literature do not apply in the present context, since the outcome we are decomposing (essentially, average test score residuals for school-teacher combinations) is a direct measure of productivity, rather than an equilibrium object. Nonetheless, research by Jackson (2013) suggests that teacher-school match components are large enough to be economically important, and that teacher mobility might be related to match quality. Thus, in this section I entertain the possibility that ϵ_{ict} contains a match component, κ_{rs}:

$$\epsilon_{ict} = \phi_{st} + \nu_{rt} + \kappa_{rs} + e_{ict} \tag{13}$$

κ_{rs} captures the possibility that teachers may be idiosyncratically more or less effective at teaching at particular schools. For example, such a match component might reflect the extent to which a teacher's teaching strengths coincide with how the principal wants lesson plans to be organized, or classrooms to be managed. The existence of the additional match component complicates interpretation of the estimated parameters, $\hat{\delta}$ and $\hat{\mu}$. In particular, a teacher's estimated quality, $\hat{\mu}$ will reflect not just her true quality, μ, but also her match component at her school (or, for transferring teachers, a weighted average of their match components at the schools at which they worked). My estimate of the variance in teaching quality within schools will now reflect both the variance in true

teaching quality, μ, and the variance in the teacher-school match component, κ_{rs}. Note, though, that the composite teacher quality estimates and the composite within school variance estimate are actually the relevant factors for examining the contribution of the current allocation of teachers to student performance inequality, since it is the combination of μ_r and κ_{rs} that determines how effective teachers are in the schools at which they are actually teaching.[9]

However, the introduction of κ_{rs} creates another mechanism by which Assumption 2 might be violated: teachers might systematically transfer to schools at which they are relatively better at teaching. Such movement toward comparative advantage would imply that mobility is not merely potentially disruptive churning, but progress toward efficient allocation of teachers to schools.

A5.2.1 The Extent of Mobility Imbalance

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

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

9. To the extent that schools and teachers can identify their potential match quality during job interviews, a model of sorting in the spirit of Lise, Meghir, and Robin (2008) or Lopes de Melo (2009) might predict that the average match component among teachers at their initial school would be positive. However, since I only compare each school's quality relative to other schools, and each teacher's quality relative to other teachers, the average match quality among schools or teachers will have no impact on the estimates. To the extent that a particular school is relatively good at identifying teachers who will be good matches during hiring, all their teachers will perform relatively well there, so this will contribute to larger school-course quality estimates, $\hat{\delta}_{sc}$, for all courses at the school. This is appropriate, since such hiring skill would be a persistent school-specific characteristic.

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

However, suppose there exists a second set of teachers of equal size that transfer from B to A because they are better matched at A than at B ($\kappa_{rB} < \kappa_{rA}$), and that the average magnitude of comparative advantage $|\kappa_{rB} - \kappa_{rA}|$ is the same across the two sets of transferrers. Then the average test score residuals at school A among the entire set of teachers who transferred between A and B will be the same as the average test score residuals at school B, so that the relative school qualities and mean teacher qualities of school A and school B will not be biased. Thus, if mobility is fully balanced, movement toward better match quality will not bias estimates of average school quality or average teacher quality across schools. Specifically, Assumption 1 will not be violated, since knowing that a teacher is a transferrer does not change the conditional expectation of ϵ_{ict} for students taught by the teacher while at the school (Assumption 1 does not condition on the direction of the transfer, so the positive conditional expectation of ϵ_{ict} for the set of teachers who transfer in will be exactly offset by the negative conditional expectation of ϵ_{ict} for the set of teachers who transfer out).

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

Fortunately, for 50.3% of schools in the sample, the fraction of transfers leaving was between .4 and .6. Furthermore, after accounting for school openings and closings and for sampling error stemming from a relatively small sample of transferring teachers associated with each school, I

show in Web Appendix Section A6 that transfer patterns are consistent with the absence systematic ladder, and inconsistent with a strong ladder.

A5.2.2 Testing for Endogenous Mobility

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

Fortunately, the two school example also suggests a possible test for mobility driven by match quality. Because the consistency of parameter estimates associated with a set of schools exhibiting balanced mobility does not require Assumption 2, if teachers are moving to better matches, a given transferring teacher should have his relatively ineffective years when teaching at the school he transferred away from and his relatively effective years when teaching at the school he transferred toward, compared to his overall average performance. Thus, for each teacher that transferred between two schools exhibiting balanced mobility, I calculate his average test score residual among students taught before transferring, and among students taught after transferring ($\bar{\hat{\epsilon}}_r^{Before}$ and $\bar{\hat{\epsilon}}_r^{After}$, respectively). The test statistic is the average difference between these residual means across all transferring teachers connecting schools featuring balanced mobility:

$$\frac{1}{|\tilde{\mathcal{R}}^B|} \sum_{r \in \tilde{\mathcal{R}}^B} (\bar{\hat{\epsilon}}_r^{After} - \bar{\hat{\epsilon}}_r^{Before}) \tag{14}$$

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

Under the null hypothesis of exogenous mobility, the average difference between these residuals across all transferring teachers should converge to 0 as the number of transferring teachers gets large. If we restrict the sample of transfers to those occurring between the 50.3% of schools whose

fraction of transfers leaving was between .4 and .6 in the data, the value of the test statistic is .0047, with a standard error of .0068. If we define the balanced schools more restrictively to be the 26.4% of schools whose fraction of transfers leaving was between .45 and .55, the test statistic is .0016, with a standard error of .0121. The point estimates suggest that teachers are on average moving toward slightly superior matches, although the magnitudes of the estimates are negligible and neither of the estimated test statistics is significant at even the 90 percent level. Notice that even if these point estimates had captured an average increase in match quality associated with a transfer, they would place an approximate upper bound on the extent of upward bias in average teacher quality estimates for the strongest net senders as a result of this type of endogenous mobility. This is because even most net senders have at least a few offsetting arriving transfers, and we would expect these teachers to have their relatively effective years at these schools, thus counteracting part of the bias.

While I fail to reject the assumption of exogenous mobility, I do not have the power to rule out that some movement is driven by match quality or other components of the error term ϵ_{ict}.[10] However, violations of exogenous mobility do not seem to be introducing significant bias into estimates of the differences in quality or average teacher quality between schools.

A6 Testing for the Existence of a Job Ladder

In order to gauge the possible bias introduced by directed mobility, I examine the extent to which teacher mobility is balanced in our data. The simplest method is to calculate the fraction of each school's associated transferring teachers who transferred out (rather than in) and examine the distribution across schools. This approach has a couple of potential drawbacks. First, when new schools are created in a district, teachers may be involuntarily reallocated by the district to the new school. Consequently, any new school in our sample will tend to have joiners make up an overwhelming fraction of their transferrers, and other schools in the district will have leavers

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

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 across schools of the fraction of transferrers who are leavers, I eliminate in-transfers to new schools in their first year, and out-transfers to that school from any other school in that year. I do the opposite for school closings.

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

Thus, I also simulate two counterfactual densities of the fraction of transferrers who are leavers at each school: one in which no job ladder exists, and a second in which a fairly strong job ladder exists. In the first case, we fix the number of transferrers at the level observed in the data for each of the 386 schools in our sample, and assume that each of those transferrers was equally likely to be a leaver or a stayer. This would be the case in the absence of a job ladder, if schools' teaching forces are remaining the same size over time. For each teacher, we take a draw, θ_r, from a Bernoulli distribution with $p = .5$, and assign this teacher to be a leaver if $\theta_r = 1$. We then calculate the fraction of each school's simulated transferrers who are leavers (denoted f_s), and sort the schools by this fraction f_s to get $\{f^1, ..., f^{386}\}$. We repeat this 100 times to get $f_b^1, ..., f_b^{386}$ for $b \in 1, ..., 100$, and average across simulated samples to get $\overline{f}^1, ..., \overline{f}^{386}$.

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

Both counterfactual densities are plotted along with the true density of f_s in Figure 4. The first thing to notice is that even with small samples of transferrers at each school, mobility is fairly balanced in the data: 50 percent of schools send between 40% and 60% of their transfers, and

78% send between 30% and 70% of their transfers. This suggests that for a large set of schools, endogenous mobility may not introduce bias into estimates of their quality and average teacher quality, relative to others in the balanced set.

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

Consequently, we turn to a second source of evidence for a job ladder, the observable characteristics of schools who are net senders or receivers of transfers. If mobility imbalance is pure small sample noise, then schools who are strong net senders should serve similar kinds of students as schools who are strong net receivers. Thus, we calculate the average student background index ($X_{ict}\beta + \tilde{Y}^i\alpha$) for schools in both the bottom quartile and top quartile of the distribution of the fraction of transferrers leaving. Strong net receivers (bottom quartile) had students who were predicted to score .07 test score standard deviations above average based only on their observable characteristics, while the strong net senders (top quartile) had students who were predicted to score .11 test score standard deviations below average. Students from the strongest net senders (bottom decile) were predicted to score .16 test score standard deviations below average. This suggests that schools serving underprivileged students are a bit more likely to lose teachers to other schools.

Overall, we see a modest amount of mobility imbalance, and while much of it is attributable to noise due to a relatively small number of transferrers at each school, mobility patterns do provide

some evidence of a job ladder, in which students serving disadvantaged students tend to be on the bottom rungs. However, most transfers seem to be driven by factors other than the academic readiness of schools' students.

WEB APPENDIX TABLES

Table A 1: TEACHER-CLASSROOM MATCH RATES BY SUBJECT

Subject	Match Rate
Algebra 1	80.8%
Algebra 2	88.6%
Biology	86.3%
Chemistry	92.6%
Civics/ELP	82.5%
English 1	83.9%
Geometry	88.0%
Physical Science	90.7%
Physics	89.3%
U.S History	86.8%

Table A 2: TEACHER-CLASSROOM MATCH RATES BY QUARTILE OF SCHOOL SIZE AND QUARTILE OF SCHOOL-AVERAGE STUDENT BACKGROUND $(X_s B + \tilde{Y}_s \alpha)$

	Match Rate	
Quartile	School Size	Avg. Background Index
Bottom 25%	87.8%	84.3%
25% to 50%	88.0%	88.1%
50% to 75%	88.9%	88.3%
Top 25%	82.4%	86.3%

Table A 3: THE PATTERN OF TEACHER MOBILITY ACROSS COURSES

Subject		Subject: Algebra 1	Algebra 2	Biology	Chemistry	E/L/P/C	English	Geometry	Physical Sciences	Physics	U.S. History
Subject	Algebra 1	5221 1.0	2227 0.427	212 0.041	43 0.008	170 0.033	245 0.047	2099 0.402	200 0.038	55 0.011	126 0.024
	Algebra 2	2227 0.756	2946 1.0	67 0.023	25 0.008	43 0.015	51 0.017	1457 0.495	63 0.021	39 0.013	38 0.013
	Biology	212 0.062	67 0.020	3433 1.0	418 0.122	194 0.057	252 0.073	79 0.023	1353 0.394	138 0.040	144 0.042
	Chemistry	43 0.032	25 0.019	418 0.314	1330 1.0	8 0.006	11 0.008	17 0.013	735 0.553	293 0.220	7 0.005
	E/L/P/C	170 0.049	43 0.012	194 0.056	8 0.002	3478 1.0	379 0.109	47 0.014	184 0.053	3 0.000	1511 0.434
	English	245 0.051	51 0.011	252 0.053	11 0.002	379 0.079	4791 1.0	65 0.014	199 0.042	2 0.000	214 0.045
	Geometry	2099 0.719	1457 0.499	79 0.027	17 0.006	47 0.016	65 0.022	2919 1.0	64 0.022	30 0.010	44 0.015
	Physical Sciences	200 0.071	63 0.022	1353 0.478	735 0.260	184 0.065	199 0.070	64 0.023	2832 1.0	388 0.137	121 0.043
	Physics	55 0.094	39 0.067	138 0.237	293 0.503	3 0.005	2 0.003	30 0.051	388 0.666	583 1.0	6 0.010
	U.S. History	126 0.047	38 0.014	144 0.053	7 0.003	1511 0.561	214 0.079	44 0.016	121 0.045	6 0.002	2693 1.0

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

WEB APPENDIX FIGURES

FIGURE A 1: DISTRIBUTIONS OF STANDARDIZED SCORES BY SUBJECT-YEAR: PART 1

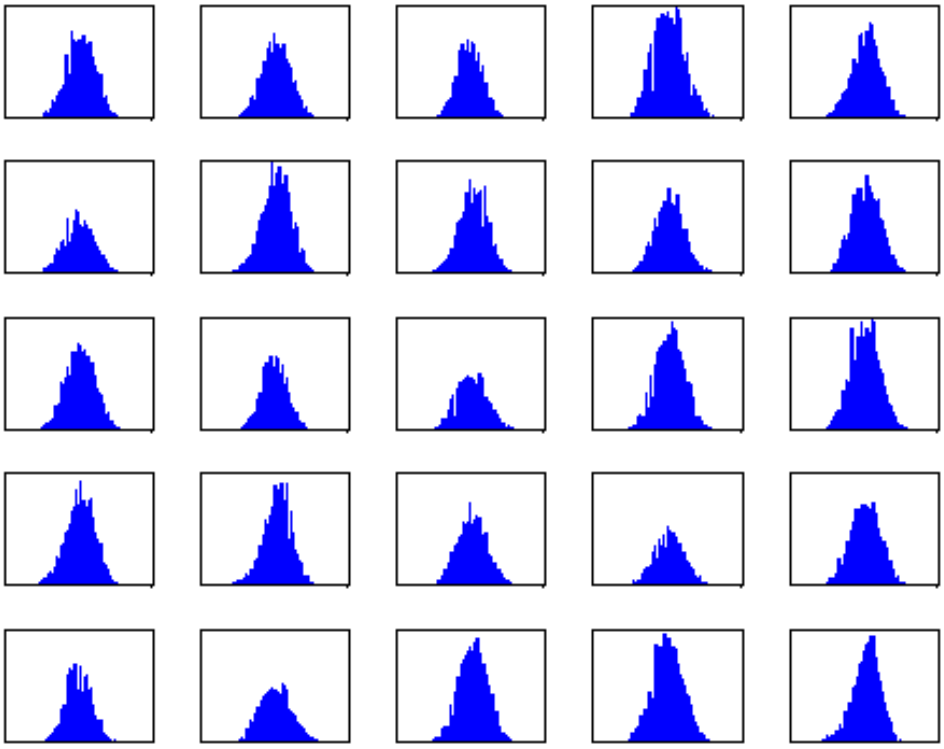

FIGURE A 2: DISTRIBUTIONS OF STANDARDIZED SCORES BY SUBJECT-YEAR: PART 2

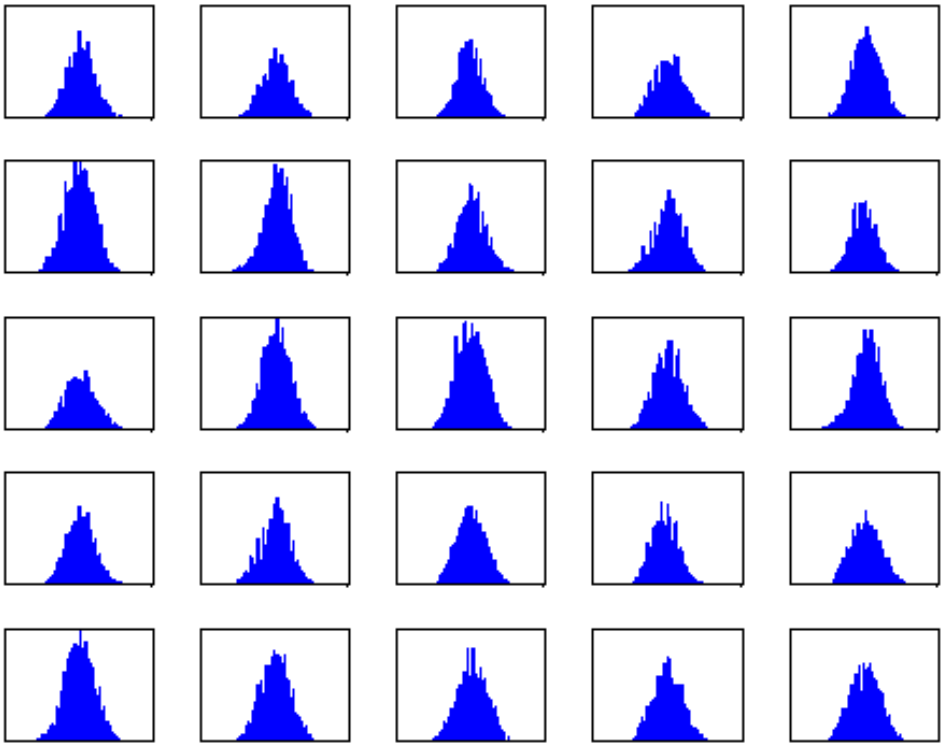

FIGURE A 3: DISTRIBUTIONS OF STANDARDIZED SCORES BY SUBJECT-YEAR: PART 3

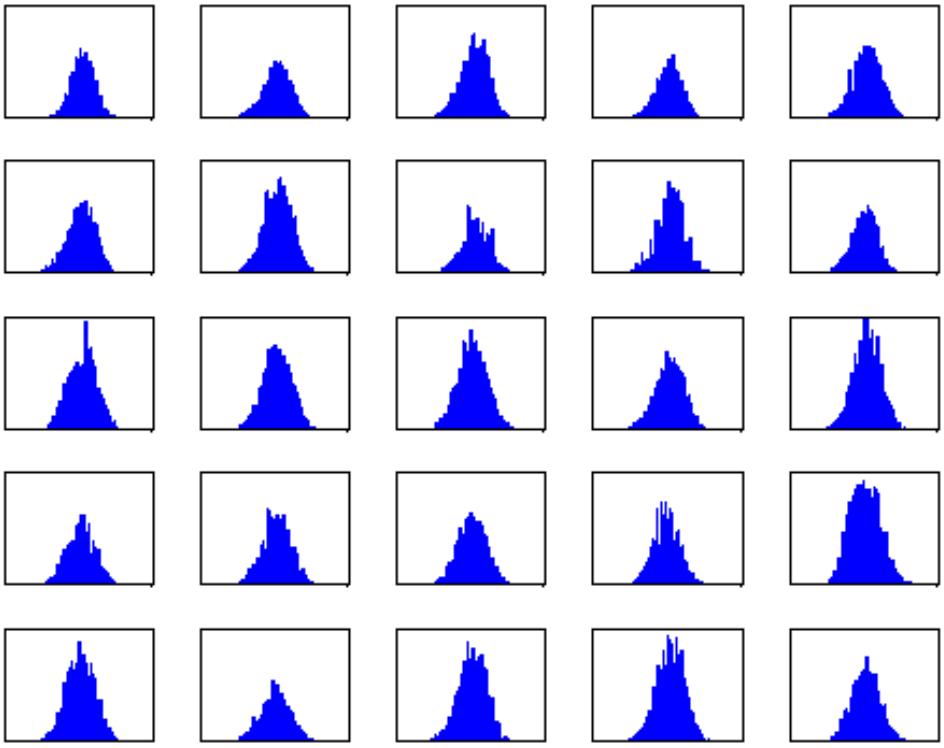

FIGURE A 4: TESTING FOR THE EXISTENCE OF A JOB LADDER: A PLOT OF THE DISTRIBUTION ACROSS SCHOOLS OF THE FRACTION OF ASSOCIATED TRANSFERRING TEACHERS THAT ARE LEAVERS (VS. ARRIVERS) USING SAMPLE DATA, SIMULATION WITH NO LADDER, AND SIMULATION WITH LADDER