Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Teacher-to-Classroom Assignment and Student Achievement

Bryan S. Graham, Geert Ridder, Petra Thiemann & Gema Zamarro

To cite this article: Bryan S. Graham, Geert Ridder, Petra Thiemann & Gema Zamarro (2023) Teacher-to-Classroom Assignment and Student Achievement, Journal of Business & Economic Statistics, 41:4, 1328-1340, DOI: 10.1080/07350015.2022.2126480

To link to this article: https://doi.org/10.1080/07350015.2022.2126480

© 2022 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 27 Oct 2022.

Submit your article to this journal

Article views: 1885

View related articles

View Crossmark data

Citing articles: 1 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ubes20

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2023, VOL. 41, NO. 4, 1328–1340
https://doi.org/10.1080/07350015.2022.2126480

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Teacher-to-Classroom Assignment and Student Achievement

Bryan S. Graham[a,b], Geert Ridder[c], Petra Thiemann[d,e], and Gema Zamarro[f,g]

[a]Department of Economics, University of California – Berkeley, Berkeley, CA; [b]National Bureau of Economic Research, Cambridge, MA; [c]Department of Economics, University of Southern California, Los Angeles, CA; [d]Department of Economics, Lund University, Lund, Sweden; [e]IZA, Bonn, Germany; [f]Department of Education Reform, University of Arkansas, Fayetteville, AR; [g]Center for Economic and Social Research (CESR), University of Southern California, Los Angeles, CA

ABSTRACT

We study the effects of counterfactual teacher-to-classroom assignments on average student achievement in U.S. elementary and middle schools. We use the Measures of Effective Teaching (MET) experiment to semiparametrically identify the average reallocation effects (AREs) of such assignments. Our identification strategy exploits the random assignment of teachers to classrooms in MET schools. To account for noncompliance of some students and teachers to the random assignment, we develop and implement a semiparametric instrumental variables estimator. We find that changes in within-district teacher assignments could have appreciable effects on student achievement. Unlike policies that aim at changing the pool of teachers (e.g., teacher tenure policies or class-size reduction measures), alternative teacher-to-classroom assignments do not require that districts hire new teachers or lay off existing ones; they raise student achievement through a more efficient deployment of *existing* teachers.

ARTICLE HISTORY
Received July 2021
Accepted August 2022

KEYWORDS
Average reallocation effects; Education production; Instrumental variables; Semiparametric methods

1. Introduction

Approximately 4 million teachers work in the public elementary and secondary education system in the United States. These teachers provide instruction to almost 50 million students, enrolled in nearly 1000 schools, across more than 13,000 school districts (McFarland et al. 2019; Snyder, de Brey, and Dillow 2017). Differences in measured student achievement are substantial across U.S. schools and across classrooms within these schools. Beginning with Hanushek (1971), a large economics of education literature attributes cross-classroom variation in student achievement to corresponding variation in (largely) latent teacher attributes. These latent attributes are referred to as teacher quality or value-added.

The implications of value-added measures (VAM) for education policy are controversial both within the academy and outside it (see Morganstein and Wasserstein 2014). The most contentious applications of VAM involve their use in teacher tenure and termination decisions (see Chetty, Friedman, and Rockoff 2012; Darling-Hammond 2015). The premise of such applications is that changes in the stock of existing teachers—specifically rooting out teachers with low VAMs and retaining those with high ones—could lead to large increases in student achievement and other life outcomes.

This article poses an entirely different question: is it possible to raise student achievement, without changes to the existing pool of teachers, by changing who teaches whom? Schools and school districts are the loci of teacher employment. To keep the analysis policy-relevant, we therefore focus on the achievement effects of different within-school and within-district teacher-to-classroom assignment policies.

For teacher assignment policies to matter, teachers must vary in their effectiveness in teaching different types of students. For example, some teachers may be especially good at teaching English language learners, minority students, or accelerated learners (see Dee 2004; Loeb, Soland, and Fox 2014). Formally educational production must be nonseparable in some teacher and student attributes (Graham, Imbens, and Ridder 2007, 2014, 2020).

We present experimental evidence of such nonseparabilities, using data from the Measures of Effective Teaching (MET) project. The MET project was conducted in six urban public school districts in the United States during two school years (2009/2010 and 2010/2011) in grades 4–10. Its goal was to evaluate different measures of teacher effectiveness; to this end, MET researchers randomly assigned teachers to classrooms within schools. We exploit the experimental design in combination with rich data on teaching practices collected throughout the experiment. Specifically, we study complementarities between (i) an observation-based, pre-experiment measure of teaching practice—Danielson's (2011) Framework for Teaching (FFT) instrument—and (ii) students' and classroom peers' baseline test scores, also measured pre-experiment. As achievement measures we use students' test score outcomes in math and English language arts (ELA).

We focus on the FFT because observational measures of classroom teaching play an important role in practice. In 98% of U.S. public schools, school principals collect observational

CONTACT Petra Thiemann petra.thiemann@nek.lu.se Department of Economics, Lund University, P.O. Box 7080, 22007 Lund, Sweden.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.
© 2022 The Authors. Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

measures and use them for a variety of purposes, for instance to determine teaching assignments, to decide upon teacher promotion and retention, and to provide feedback to teachers (U.S. Department of Education 2020). Such measures are also increasingly used in research (e.g., Garrett and Steinberg 2015; Araujo et al. 2016; Burgess, Rawal, and Taylor 2021; notably, Aucejo et al. 2022, investigate interaction effects between the FFT and classroom composition in the MET data). Because of their practical relevance and wide applicability, observational measures are a natural starting point for the study of alternative teacher-to-classroom assignments.

To quantify the potential achievement effects of alternative teacher-to-classroom assignments, we estimate the hypothetical test score gains associated with a reallocation that maximizes average achievement. Because our model allows for both complementarities and nonlinearity in the underlying variables (FFT and student baseline achievement), determining an outcome-maximizing assignment is nontrivial. We compute the optimal assignment using linear programming methods under the constraint that the existing pool of teachers remains unchanged (see Bhattacharya 2009).

We find that within-district changes in teacher-to-classroom assignments could increase average classroom student achievement by as much as 0.04 standard deviations. This effect corresponds to the estimated difference in average achievement across the teacher-to-classroom assignment which maximizes aggregate achievement versus the one which minimizes it. Perhaps more realistically, comparing the status quo assignment in MET schools—which was generated by random assignment of teachers to classrooms—with the optimal one generates an estimated increase in average test scores of 0.02 standard deviations.

To benchmark these effect sizes consider a policy which removes the bottom $\tau \times 100\%$ of teachers from classrooms—sorted according to their VAM—and replaces them with average teachers (i.e., teachers with VAMs of zero). Assuming a Gaussian distribution for teacher value-added, the effect of such an intervention would be to increase the mean of the student test score distribution by

$$(1-\tau)\,\sigma\,\frac{\phi\left(\frac{q_\tau}{\sigma}\right)}{1-\Phi\left(\frac{q_\tau}{\sigma}\right)}$$

standard deviations. Here σ corresponds to the standard deviation of teacher value-added and q_τ to its τth quantile. Rockoff (2004, Table 2) and Rothstein (2010, Table 6) estimate a standard deviation of teacher value-added of between 0.10 and 0.15. Taking the larger estimate and setting $\tau = 0.05$ (0.10) generates an expected increase in student test scores of 0.015 (0.026) standard deviations; this is comparable to our reallocation effects. Replacing 5% (10%) of teachers would be difficult to do in practice. It would be even more difficult to correctly identify the bottom 5% (10%) of teachers (according to VA) and replace them with average ones. We conclude that the achievement effects of teacher assignment policies are meaningful.

In contrast to policies which replace or hire additional teachers, the within-school and within-district teacher reassignment policies we explore in this article do not require that districts lay off existing teachers or attract new ones. Of course, reassigning teachers across classrooms, and especially across schools within a district, may involve other types of costs. For example, many school districts operate under collective bargaining agreements which give senior teachers partial control over their school assignment (e.g., Cohen-Vogel, Feng, and Osborne-Lampkin 2013).

Work by Susanna Loeb and coauthors suggests that U.S. school districts have *de facto* teacher-to-classroom assignment policies (Kalogrides, Loeb, and Beteille 2011; Grissom, Kalogrides, and Loeb 2015). For example, they find that less experienced, minority, and female teachers are more likely to be assigned to predominantly minority classrooms. They also present evidence that principals use teacher assignments as mechanisms for retaining teachers—as well as for encouraging less effective teachers to leave—and that more experienced teachers exert more influence on classroom assignment decisions. The present article helps researchers and policy-makers understand the achievement effects of such policies and the potential benefits of alternative ones. The findings presented below suggest that teacher-to-classroom assignment policies are consequential and that changes to them could meaningfully increase average student achievement.

In addition to our substantive results, we present new identification results for average reallocation effects (AREs). Identification and estimation of AREs under (conditional) exogeneity is considered by Graham, Imbens, and Ridder (2014, 2020). These results do not apply directly here. Although teachers were randomly assigned to classrooms as part of the MET experiment, compliance was imperfect. Furthermore some students moved across classrooms after the random assignment of teachers, which raises concerns about bias due to endogenous student sorting. We develop a semiparametric instrumental variables estimator (e.g., Ai and Chen 2003) which corrects for student and teacher noncompliance. Our analysis highlights how complex identification can be in the context of multi-population matching models where agents sort endogenously.

This article leaves several important questions for further research. First, assignments based upon a *different* combination of teacher and classroom attributes could lead to even greater achievement gains. Second, one may consider objective functions that do no focus on average test scores but instead on test score gaps or proficiency levels. Third, outcomes other than math and ELA achievement (e.g., socio-emotional skills) may also be of interest. Finally, we abstract from issues that will likely arise when an allocation policy is adopted in practice. For instance, implementing an optimal allocation could generate general equilibrium effects or teacher noncompliance patterns, which we cannot predict in the present study (see Carrell, Sacerdote, and West 2013). We leave explorations along these lines to future work and consider this article as a first pass that establishes the feasibility of recovering match effects from imperfect experimental data and shows that the resulting reallocation effects that depend upon these match effects can be substantial.

2. Model and Identification

Our goal is to identify the average achievement effects of alternative assignments of teachers to MET classrooms. These are average reallocation effects (AREs), as introduced by Graham,

Imbens, and Ridder (2007, 2014). The identification challenge is to use the observed MET teacher-to-classroom assignments and outcomes to recover these AREs.

Our analysis is based upon experimentally generated combinations of student and teacher attributes, that is, it exploits the random assignment of teachers to classrooms within schools in the MET experiment. Like in many other field experiments, various deviations from MET's intended protocol complicate the analysis. In this section we outline a semiparametric model of educational production and consider its identification based upon the MET project *as implemented*, using the MET data *as collected*.

It is useful, however, to first explore nonparametric identification of reallocation effects under an ideal implementation of the MET project (henceforth MET *as designed*). Such an approach clarifies how the extra restrictions introduced below allow for the identification of reallocation effects despite noncompliance, attrition, and other deviations from the intended experimental protocol.

2.1. Nonparametric Identification Under Ideal Circumstances

The setting features two populations, one of students and the other of teachers. Each student is distinguished by an observed attribute X_i, in our case a measure of baseline academic achievement, and an unobserved attribute, say "student ability," V_i (shorthand for latent student attributes associated with higher test scores). Similarly, each teacher is characterized by an observed attribute W_i, in our case an observation-based measure of teaching practice, and an unobserved attribute, say "teacher quality," U_i (shorthand for latent teacher attributes associated with higher test scores).

Let $i = 1, \ldots, N$ index students. Let C be the total number of MET classrooms or equivalently teachers. We define G_i to be a $C \times 1$ vector of classroom assignment indicators. The cth element of G_i equals one if student i is in classroom $c \in \{1, \ldots, C\}$ and zero otherwise. The indices of student i's peers or classmates are therefore given by the index set

$$p(i) = \left\{ j : G_i = G_j,\ i \neq j \right\}.$$

Next we define the peer average attribute as $\bar{X}_{p(i)} = \frac{1}{|p(i)|} \sum_{j \in p(i)} X_j$ (i.e., the average of the characteristic X across student i's peers). We define $\bar{V}_{p(i)}$ similarly.

The MET project protocol did not impose any requirements on how students, in a given school-by-grade cell, were divided into classrooms. Evidently schools followed their existing procedures for dividing students within a grade into separate classrooms. An implication of this observation is that the MET experiment implies no restrictions on the joint density

$$f_{X_i, V_i, \bar{X}_{p(i)}, \bar{V}_{p(i)}}(x, v, \bar{x}, \bar{v}), \tag{1}$$

beyond the requirement that the density be feasible. For example, if most schools tracked students by prior test scores, then we would expect X_i and $\bar{X}_{p(i)}$ to positively covary. If, instead, students were randomly assigned to classrooms and hence peers, we would have, ignoring finite population issues, the factorization

$$f_{X_i, V_i, \bar{X}_{p(i)}, \bar{V}_{p(i)}}(x, v, \bar{x}, \bar{v}) = f_{X_i, V_i}(x, v) f_{\bar{X}_{p(i)}, \bar{V}_{p(i)}}(\bar{x}, \bar{v}).$$

Our analysis allows for arbitrary dependence between own and peer attributes, both observed and unobserved, and consequently is agnostic regarding the protocol used to group students into classrooms.

Two implications of this agnosticism are (i) our analysis is necessarily silent about the presence and nature of any peer group effects, and (ii) it is likely that more complicated policies, involving *simultaneously* regrouping students into new classes *and* reassigning teachers to them, could raise achievement by more than what is feasible via reassignments of teachers to *existing* classrooms alone, which is the class of policies we consider. Learning about the effects of policies which simultaneously regroup students and reassign teachers would require double randomization (see Graham 2008; Graham, Imbens, and Ridder 2010, 2020).

Although nothing about the MET protocol generates restrictions on the joint density (1), random assignment of teachers to classrooms—however, formed—ensures that

$$\begin{aligned} &f_{X_i, V_i, \bar{X}_{p(i)}, \bar{V}_{p(i)}, W_i, U_i}(x, v, \bar{x}, \bar{v}, w, u) \\ &\quad = f_{X_i, V_i, \bar{X}_{p(i)}, \bar{V}_{p(i)}}(x, v, \bar{x}, \bar{v}) f_{W_i, U_i}(w, u). \end{aligned} \tag{2}$$

Here W_i and U_i denote the observed and unobserved attributes of the teacher assigned to the classroom of student i. A perfect implementation of MET as designed would ensure that student and teacher attributes vary independently of each other. Our research design is fundamentally based upon restriction (2).

Let Y_i be an end-of-year measure of student achievement, generated according to

$$Y_i = g\left(X_i, \bar{X}_{p(i)}, W_i, V_i, \bar{V}_{p(i)}, U_i\right). \tag{3}$$

Other than the restriction that observed and unobserved peer attributes enter as means, Equation (3) imposes no restrictions on educational production. Under restriction (2) the conditional mean of the outcome given observed own, peer, and teacher attributes equals

$$\begin{aligned} &\mathbb{E}\left[Y_i \middle| X_i = x, \bar{X}_{p(i)} = \bar{x}, W_i = w \right] \\ &\quad = \iiint \left[g(x, \bar{x}, w, v, \bar{v}, u) f_{V_i, \bar{V}_{p(i)} | X_i, X_{p(i)}}(v, \bar{v} | x, \bar{x}) \right. \\ &\qquad \left. \times f_{U_i | W_i}(u | w) \right] \mathrm{d}v \mathrm{d}\bar{v} \mathrm{d}u \\ &\quad = m^{\mathrm{amf}}(x, \bar{x}, w). \end{aligned} \tag{4}$$

Equation (4) coincides with (a variant of) the Average Match Function (AMF) estimand discussed by Graham, Imbens, and Ridder (2014, 2020). The AMF can be used to identify AREs. Our setting—which involves multiple students being matched to a single teacher—is somewhat more complicated than the one-to-one matching settings considered by Graham, Imbens, and Ridder (2014, 2020). One solution to this "problem" would be to average Equation (3) across all students in the same classroom and work directly with those averages. As will become apparent below, however, working with a student-level model makes it

Table 1. Feasible teacher reassignments.

Classroom type	Status quo $\Pr(W_i=1\mid X_i,\bar{X}_{p(i)})$	Counterfactual $\Pr(\tilde{W}_i=1\mid X_i,\bar{X}_{p(i)})$	X_i	$\bar{X}_{p(i)}$	$f(X_i,\bar{X}_{p(i)})$
000	$\frac{1}{2}$	$\frac{1}{3}$	0	0	$\frac{1}{4}$
			0	0	$\frac{1}{4}$
			0	0	$\frac{1}{4}$
001	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{6}$
			0	$\frac{1}{2}$	$\frac{1}{6}$
			1	0	$\frac{1}{12}$
011	$\frac{1}{2}$	$\frac{1}{2}$	0	1	$\frac{1}{12}$
			1	$\frac{1}{2}$	$\frac{1}{6}$
			1	$\frac{1}{2}$	$\frac{1}{6}$
111	$\frac{1}{2}$	$\frac{2}{3}$	1	1	$\frac{1}{4}$
			1	1	$\frac{1}{4}$
			1	1	$\frac{1}{4}$

Note: The population fraction of type $X_i = 1$ students is $\frac{1}{2}$ and that of type $W_i = 1$ teachers is also $\frac{1}{2}$. Classrooms of three students each are formed, such that the frequency of each of the four possible classroom configurations is $\frac{1}{4}$ in the population of classrooms of size 3. Under the status quo teachers are assigned to classrooms at random; in the counterfactual teachers are assigned more assortatively. See the main text for more information.

easier to deal with noncompliance and attrition, which have distinctly student-level features. It also connects our results more directly with existing empirical work in the economics of K-to-12 education, where student-level modeling predominates, and results in greater statistical power.

The decision to model outcomes at the student level makes the analysis of teacher reassignments a bit more complicated, at least superficially. To clarify the issues involved it is helpful to consider an extended example. Assume there are two *types* of students, $X_i \in \{0,1\}$, and two *types* of teachers, $W_i \in \{0,1\}$. For simplicity assume that the population fractions of type $X_i = 1$ students and type $W_i = 1$ teachers both equal one-half, that is, half of the students are of type 1 and half of the students are taught by a teacher of type 1. Assume, again to keep things simple, that classrooms consist of three students each.

Table 1 summarizes this basic set-up. Column 1 lists classroom types. For example, a 000 classroom consists of all type-0 students. There are four possible classroom types, each assumed to occur with a frequency of one-fourth. The status quo mechanism for grouping students into classrooms induces a joint distribution of own and peer average attributes. This joint distribution is given in the right-most column of Table 1. For instance, $\frac{1}{4}$ (3 out of 12) of the students are in a classroom with two type-0 peers, so that $f_{X_i,\bar{X}_{p(i)}}(0,0) = \frac{1}{4}$, and $\frac{1}{6}$ (2 out of 12) of the students are in a classroom with one type-0 and one type-1 peer, so that $f_{X_i,\bar{X}_{p(i)}}(0,\frac{1}{2}) = \frac{1}{6}$. The MET experiment implies no restrictions on the joint density $f_{X_i,\bar{X}_{p(i)}}(x,\bar{x})$, consequently we only consider policies which leave it unchanged.

Next assume, as was the case in the MET experiment, that under the status quo teachers are randomly assigned to classrooms. This induces the conditional distribution of W_i given X_i and $\bar{X}_{p(i)}$ reported in column 2 of Table 1. Of course, from this conditional distribution, and the marginal for X_i and $\bar{X}_{p(i)}$, we can recover the joint distribution of own type, peer average type, and teacher type (i.e., of X_i, $\bar{X}_{p(i)}$ and W_i).

Now consider the AMF: $m^{\text{amf}}(x,\bar{x},w)$. Consider the subpopulation of students with $X_i = 1$ and $\bar{X}_{p(i)} = \frac{1}{2}$. Inspecting Table 1, this subpopulation represents $\frac{1}{6}$ of all students (right-most column of Table 1). If we assign to students in this subpopulation a teacher of type $W_i = 1$, then the expected outcome coincides with $m^{\text{amf}}\left(1,\frac{1}{2},1\right)$. Under random assignment of teachers the probability of assigning a type-1 teacher is the same for all subpopulations of students.

Finally consider a counterfactual assignment of teachers to classrooms. Since we leave the composition of classrooms unchanged, $f_{X_i,\bar{X}_{p(i)}}(x,\bar{x})$ is left unmodified. The counterfactual assignment therefore corresponds to a conditional distribution for teacher type, $\tilde{f}_{\tilde{W}_i\mid X_i,\bar{X}_{p(i)}}(w\mid x,\bar{x})$ which satisfies the feasibility condition:

$$\iint \tilde{f}_{\tilde{W}_i\mid X_i,\bar{X}_{p(i)}}(w\mid x,\bar{x}) f_{X_i,\bar{X}_{p(i)}}(x,\bar{x})\,\mathrm{d}x\mathrm{d}\bar{x} = f(w) \tag{5}$$

for all $w \in \mathbb{W}$. Here $\tilde{f}$ denotes a counterfactual distribution, while f denotes a status quo one. We use $\tilde{W}_i$ to denote an assignment from the counterfactual distribution. Note that by feasibility of an assignment $\tilde{W}_i \overset{D}{=} W_i$ marginally, but will differ conditional on student attributes. Condition (5), as discussed by Graham, Imbens, and Ridder (2014), allows for degenerate conditional distributions, as might occur under a perfectly positive assortative matching.

Average achievement under a counterfactual teacher-to-classroom assignment equals:

$$\beta^{\text{are}}\left(\tilde{f}\right) = \iint \left[\int m^{\text{amf}}(x,\bar{x},w)\tilde{f}_{\tilde{W}_i\mid X_i,\bar{X}_{p(i)}}(w\mid x,\bar{x})\,\mathrm{d}w\right] f_{X_i,\bar{X}_{p(i)}}(x,\bar{x})\,\mathrm{d}x\mathrm{d}\bar{x}. \tag{6}$$

Since all the terms to the right of the equality are identified, so too is the ARE. Conceptually we first—see the inner integral in Equation (6)—compute the expected outcome in each type of classroom (e.g., $X_i = x$ and $\bar{X}_{p(i)} = \bar{x}$) given its new teacher assignment (e.g., to type $\tilde{W}_i = w$). We then—see the outer two integrals in Equation (6)—average over the status quo distribution of $X_i, \bar{X}_{p(i)}$, which is left unchanged. This yields average student achievement under the new assignment of teachers to classrooms.

In addition to the feasibility condition (5) we need to also rule out allocations that assign different teachers to students in the same classrooms. Note that $m^{\text{amf}}(x,\bar{x},w)$ is the average outcome for the subpopulation of students of type $X_i = x$ with peers $\bar{X}_{p(i)} = \bar{x}$. For example, in Table 1 classroom 001 has students from two subpopulations so defined. Assignment of teachers to subpopulations of students opens up the possibility that a classroom is assigned to teachers of different types for its constituent subgroups of students. If, as indicated in Table 1, the teacher-type assignment probability is the same for all subpopulations of students represented in a classroom, then the ARE in Equation (6) coincides with one based on direct assignment of teachers to classrooms. This implicit restriction on teacher assignments

provides a link between models for individual outcomes and classroom-level reallocations.

2.2. Semiparametric Identification Under MET as Implemented

In the MET experiment as implemented not all teachers and students appear in their assigned classrooms. This occurs both due to attrition (e.g., when a student changes schools prior to follow-up) and due to actual noncompliance (e.g., when a teacher teaches in a classroom different from their randomly assigned one).

This section describes our approach to identifying AREs in MET as implemented. Relative to the previous section we impose two types of additional restrictions. First, we work with a semiparametric, as opposed to a nonparametric, educational production function. Second, we make behavioural assumptions regarding the nature of noncompliance. Both sets of assumptions are (partially) testable.

2.2.1. Educational Production Function

Our first set of restrictions involve the form of the educational production function. A key restriction we impose is that *unobserved* student, peer, and teacher attributes enter separably. Although this assumption features in the majority of economics of education empirical work (e.g., Chetty, Friedman, and Rockoff 2014), it is restrictive. We also discretize the observed student and teacher attributes. This allows us to work with a parsimonously parameterized educational production function that nevertheless accommodates complex patterns of complementarity between student and teacher attributes. Discretization also allows us to apply linear programming methods to study counterfactual assignments (see Graham, Imbens, and Ridder 2007; Bhattacharya 2009).

Specifically we let X_i be a vector of indicators for each of K "types" of students. Types correspond to intervals of baseline test scores. Our preferred specification works with $K = 3$ types of students: those with low, medium, and high baseline test scores. In this case X_i is a 2×1 vector of dummies for whether student i's baseline test score was in the medium or high range (with the low range being the omitted group). This definition of X_i means that $\bar{X}_{p(i)}$ equals the 2×1 vector of fractions of peers in the medium and high baseline categories (with the fraction low range omitted).

We discretize the distribution of the teacher attribute similarly: W_i is a vector of indicators for L different ranges of FFT scores. In our preferred specification we also work with $L = 3$ types of teachers: those with low, medium, and high FFT scores. Hence, W_i is again a 2×1 vector of dummies for whether the teacher of student i's FFT score was in the medium or high range (with the low range again being the omitted group).

We assess the sensitivity of our results to coarser and finer discretizations of the baseline test score and FFT distributions. Specifically we look at $K = L = 2$ and $K = L = 4$ discretizations (see Section B.2 in the supplementary materials).

We posit that end-of-school year achievement for student i is generated according to

$$Y_i = \underbrace{\alpha + X_i'\beta + V_i}_{\text{Student Ability}} + \underbrace{\bar{X}_{p(i)}'\gamma + \rho \bar{V}_{p(i)}}_{\text{Peer Effect}} + \underbrace{W_i'\delta + U_i}_{\text{Teacher Quality}} + \underbrace{\left(X_i \otimes \bar{X}_{p(i)}\right)'\zeta}_{\text{Student-Peer Complementarity}} + \underbrace{(X_i \otimes W_i)'\eta + \left(W_i \otimes \bar{X}_{p(i)}\right)'\lambda}_{\text{Student-Teacher Complementarity}} \tag{7}$$

Observe that—as noted above—own, V_i, peer, $\bar{V}_{p(i)}$, and teacher, U_i, unobservables enter Equation (7) additively. The labeled grouping of terms in Equation (7) highlights the flexibility of our model relative to those typically employed by researchers. As in traditional models, observed and unobserved student and teacher attributes are posited to directly influence achievement. We add to this standard set-up the possibility of complementarity between own and peer attributes, and complementarity between own and teacher attributes. Additionally our model allows for both observed and *unobserved* peer attributes to influence achievement.

Conditional on working with a discrete student and teacher type space, Equation (7) is unrestrictive in how own and teacher attributes interact to generate achievement. In contrast, Equation (7) restricts the effect of peers' observed composition on the outcome. Partition $\zeta = (\zeta_1, \ldots, \zeta_{K-1})$ and similarly partition $\lambda = (\lambda_1, \ldots, \lambda_{L-1})$. The $(K-1) \times 1$ gradient of student i's outcome with respect to peer composition is

$$\frac{\partial Y_i}{\partial \bar{X}_{p(i)}} = \gamma + \sum_{k=1}^{K-1} X_{ki}\zeta_k + \sum_{l=1}^{L-1} W_{li}\lambda_l, \tag{8}$$

which is constant in $\bar{X}_{p(i)}$, although varying heterogenously with student and teacher type. Put differently convexity/concavity in $\bar{X}_{p(i)}$ is ruled out by Equation (7). It should be noted that the MET data, in which the assignment of peers is not random, are not suitable for estimating peer effects (nonlinear or otherwise).

For completeness we also include the interaction of teacher type with peer composition—the $\left(W_i \otimes \bar{X}_{p(i)}\right)$ regressor in Equation (7)—although λ is poorly identified in practice. Due to our limited sample size, we do not include the third order interactions of own, peer, and teacher types.

Relative to a standard "linear-in-means" type model typically fitted to datasets like ours (e.g., Hanushek, Kain, and Rivkin 2004):

$$Y_i = \alpha + X_i'\beta + \bar{X}_{p(i)}'\gamma + W_i'\delta + V_i + U_i, \tag{9}$$

Equation (7) is rather flexible. It allows for rich interactions in observed own, peer, and teacher attributes and is explicit in that both observed *and* unobserved peer attributes may influence own achievement. The "linear-in-means" model (9) presumes homogenous effects and does not explicitly incorporate unobserved peer attributes (for models with unobserved peer attributes, see Manski 1993; Graham 2008).

As mentioned above, a student's assigned teacher and peers may deviate from her realized ones due to attrition and noncompliance. To coherently discuss our assumptions about these issues we require some additional notation. Let W_i^* and $\bar{X}_{p^*(i)}$ denote student i's *assigned* teacher and peer attribute (here $p^*(i)$ is the index set of i's *assigned* classmates). Random assignment of teachers to classrooms ensures that a student's *assigned*

teacher's attributes are independent of her own unobservables:

$$\mathbb{E}\left[\, V_i \middle|\, X_i, \bar{X}_{p^*(i)}, W_i^* \right] = \mathbb{E}\left[\, V_i \middle|\, X_i, \bar{X}_{p^*(i)} \right] \stackrel{\text{def}}{\equiv} g_1\left(X_i, \bar{X}_{p^*(i)}\right). \tag{10}$$

Here $g_1(x, \bar{x})$ is unrestricted. Under double randomization, with students additionally grouped into classes at random, we would have the further restriction

$$\mathbb{E}\left[\, V_i \middle|\, X_i, \bar{X}_{p^*(i)} \right] = \mathbb{E}\left[\, V_i \middle|\, X_i \right].$$

However, since the MET experiment placed no restrictions on how students were grouped into classrooms, we cannot rule out the possibility that a student's peer characteristics, $\bar{X}_{p^*(i)}$, predict her own unobserved ability, V_i. Consequently our data are necessarily silent about the presence and nature of any peer group effects in learning. This limitation does not limit our ability to study the effects of teacher reallocations, because we leave the student composition of classrooms—and hence the "peer effect"—fixed in our counterfactual experiments.

Finally, even with double randomization, we would still have $\mathbb{E}\left[\, V_i \middle|\, X_i \right] \neq 0$. Observed and unobserved attributes may naturally covary in any population (for example, average hours of sleep, which is latent in our setting, plausibly covaries with baseline achievement and also influences the outcome). Such covariance is only a problem if, as is true in the traditional program evaluation setting, the policies of interest induce changes in the marginal distribution of X_i—and hence the joint distribution of X_i and V_i. This is not the case here: any reallocations leave the joint distribution of X_i and V_i unchanged.

The MET protocol also ensures that assigned peer unobservables, $\bar{V}_{p^*(i)}$, are independent of the observed attributes of one's assigned teacher:

$$\mathbb{E}\left[\bar{V}_{p^*(i)} \middle| X_i, \bar{X}_{p^*(i)}, W_i^* \right] = \mathbb{E}\left[\bar{V}_{p^*(i)} \middle| X_i, \bar{X}_{p^*(i)} \right] \stackrel{\text{def}}{\equiv} g_2\left(X_i, \bar{X}_{p^*(i)}\right), \tag{11}$$

with $g_2\left(X_i, \bar{X}_{p^*(i)}\right)$, like $g_1\left(X_i, \bar{X}_{p^*(i)}\right)$, unrestricted.

Random assignment of teachers to classrooms also ensures independence of the unobserved attribute of a student's *assigned* teacher and observed student and peer characteristics:

$$\mathbb{E}\left[\, U_i^* \middle|\, X_i, \bar{X}_{p^*(i)}, W_i^* \right] = \mathbb{E}\left[\, U_i^* \middle|\, W_i^* \right] = 0. \tag{12}$$

The second equality is a normalization; reallocations leave the joint distribution of U_i and W_i unchanged, so we are free to normalize this mean to zero.

Under MET *as designed* we could identify AREs using Equations (10)–(12). To see this let, as would be true under perfect compliance $W_i = W_i^*$ and $\bar{X}_{p(i)} = \bar{X}_{p^*(i)}$ for all $i = 1, \ldots, N$. Using Equations (10)–(12) in combination with the education production function outlined in Equation (7) yields, after some algebraic manipulation, the partially linear regression model (e.g., Robinson 1988):

$$Y_i = W_i'\delta + (X_i \otimes W_i)'\,\eta + \left(W_i \otimes \bar{X}_{p(i)}\right)'\lambda + h\left(X_i, \bar{X}_{p(i)}\right) + A_i \tag{13}$$

with $\mathbb{E}\left[A_i \middle|\, X_i, \bar{X}_{p(i)}, W_i \right] = 0$ for

$$A_i \stackrel{\text{def}}{\equiv} \left[V_i - g_1\left(X_i, \bar{X}_{p(i)}\right)\right] + \rho\left[\bar{V}_{p(i)} - g_2\left(X_i, \bar{X}_{p(i)}\right)\right] + U_i, \tag{14}$$

and where the nonparametric regression component equals

$$h\left(X_i, \bar{X}_{p(i)}\right) \stackrel{\text{def}}{\equiv} \alpha + X_i'\beta + g_1\left(X_i, \bar{X}_{p(i)}\right) + \bar{X}_{p(i)}'\gamma + \rho g_2\left(X_i, \bar{X}_{p(i)}\right) + \left(X_i \otimes \bar{X}_{p(i)}\right)'\zeta. \tag{15}$$

Note, even under this perfect experiment, we cannot identify β, γ, and ζ; these terms are confounded by $g_1\left(X_i, \bar{X}_{p(i)}\right)$ and $g_2\left(X_i, \bar{X}_{p(i)}\right)$ and hence absorbed into the nonparametric component of the regression model. This lack of identification reflects the inherent inability of the MET experiment to tell us anything about peer group effects. We also cannot disentangle the teacher-student complementarity, η, from the teacher-peer complementarity, λ, because classroom composition is nonrandom. Nevertheless, knowledge of δ, η, and λ is sufficient to identify the class of reallocation effects we focus upon, because the reallocations we consider leave the joint distribution of student and peer characteristics unchanged.

2.2.2. Patterns of Noncompliance

Unfortunately, we do not observe student outcomes under full compliance. Noncompliance may induce correlation between A_i and X_i, $\bar{X}_{p(i)}$ and W_i in regression model (13). Our solution to this problem is to construct instrumental variables for *observed* teacher and peer attributes, W_i and $\bar{X}_{p(i)}$—which necessarily reflect any noncompliance and attrition on the part of teachers and students—from the *assigned* values, W_i^* and $\bar{X}_{p^*(i)}$.

Rigorously justifying this approach requires imposing restrictions on how, for example, realized and assigned teacher quality relate to one another. In other words, we need to exclude systematic patterns of student or teacher switching behavior based on the characteristics of the *assigned* teachers or *assigned* peers that could bias our IV estimates.

Assumption 1. (Idiosyncratic Teacher Deviations)

$$\mathbb{E}\left[\, U_i - U_i^* \middle|\, X_i, \bar{X}_{p^*(i)}, W_i^* \right] = 0. \tag{16}$$

Assumption 1 implies that the *difference* between *realized* and *assigned* (unobserved) teacher quality cannot be predicted by own and assigned peer and teacher observables. The assumption can be violated, for instance, if students who are assigned a low-FFT teacher systematically move into classrooms with a teacher that has a higher unobserved quality, U_i, than their assigned teacher. In this case, estimates of teacher FFT using assigned teacher FFT as instrument could be biased.

While Assumption 1 is not directly testable, we can perform the following plausibility test. Let $R_i - R_i^*$ be the difference between the realized and assigned value of some observed teacher attribute other than W_i (e.g., years of teaching experience). Under Equation (16), if we compute the OLS fit of this difference onto 1, X_i, W_i^*, and $\bar{X}_{p^*(i)}$, a test for the joint significance of the nonconstant regressors should accept the null of no effect. Finding that, for example, students *assigned* to classrooms with a low-FFT teacher tend to *move into* classrooms with more experienced teachers suggests that Assumption 1 may be implausible.

Assumption 1 and Equation (12) yield the mean independence restriction

$$\mathbb{E}\left[\, U_i \middle|\, X_i, W_i^*, \bar{X}_{p^*(i)} \right] = 0. \tag{17}$$

This equation imposes restrictions on the unobserved attribute of student i's *realized* teacher. It is this latent variable which drives the student outcome actually observed.

Our second assumption involves the relationship between the unobserved attributes of a student's assigned peers and those of her realized peers. These two variables will differ if some students switch out of their assigned classrooms.

Assumption 2. (Conditionally Idiosyncratic Peer Deviations)

$$\mathbb{E}\left[\bar{V}_{p(i)}-\bar{V}_{p^*(i)}\middle|\, X_i,\bar{X}_{p^*(i)},W_i^*\right]=\mathbb{E}\left[\bar{V}_{p(i)}-\bar{V}_{p^*(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]. \tag{18}$$

Assumption 2 implies that the *difference* between *realized* and *assigned* unobserved peer quality cannot be predicted by assigned teacher observables. We do allow for these deviations to covary with a student's type and the assigned composition of her peers. Assumption 2 can be violated, for instance, if assigned peers with high unobserved quality, $V_{p^*(i)}$, move out of the classroom if assigned to a low-FFT teacher. In this case assignment to a low-FFT teacher could affect test scores via changes in classroom composition, thus, potentially creating a bias in the IV estimates of teacher FFT.

We can assess the plausibility of Assumption 2 using observed peer attributes. Finding, for example, that—conditional on own type, X_i, and assigned peers' average type, $\bar{X}^*_{p(i)}$—assigned teacher quality, W_i^*, predicts differences between the realized and assigned values of observed peer attributes (e.g., gender, race) provides evidence against Assumption 2.

Assumption 2 and Equation (11) yield a second mean independence restriction of

$$\mathbb{E}\left[\,\bar{V}_{p(i)}\middle|\, W_i^*,X_i,\bar{X}_{p^*(i)}\right]=g_2^*\left(X_i,\bar{X}_{p^*(i)}\right), \tag{19}$$

where $g_2^*\left(X_i,\bar{X}_{p^*(i)}\right)\overset{\text{def}}{\equiv}\mathbb{E}\left[\,\bar{V}_{p(i)}-\bar{V}_{p^*(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]+g_2\left(X_i,\bar{X}_{p^*(i)}\right)$ is unrestricted.

The experiment-generated restrictions-Equations (10)–(12)—in conjunction with our two (informally testable) assumptions about deviations from the experiment protocol—Assumptions 1 and 2—together imply the following conditional moment restriction:

$$\begin{aligned}&\mathbb{E}\left[\,U_i+V_i+\rho\bar{V}_{p(i)}\middle|\, W_i^*,X_i,\bar{X}_{p^*(i)}\right]\\&\quad=g_1\left(X_i,\bar{X}_{p^*(i)}\right)+\rho g_2^*\left(X_i,\bar{X}_{p^*(i)}\right).\end{aligned} \tag{20}$$

We wish to emphasize two features of restriction (20). First, the conditioning variables are *assigned* peer and teacher attributes, not their realized counterparts. This reflects our strategy of using assignment constructs as instruments. Second, any function of W_i^*, as well as interactions of such functions with functions of X_i and $\bar{X}_{p^*(i)}$ do not predict the composite error $U_i+V_i+\rho\bar{V}_{p(i)}$ conditional on X_i and $\bar{X}_{p^*(i)}$; hence, such terms are valid instrumental variables.

More specifically we redefine h to equal

$$\begin{aligned}h\left(X_i,\bar{X}_{p^*(i)},\bar{X}_{p(i)}\right)\overset{\text{def}}{\equiv}&\alpha+X_i'\beta+g_1\left(X_i,\bar{X}_{p^*(i)}\right)+\bar{X}_{p(i)}'\gamma\\&+\rho g_2^*\left(X_i,\bar{X}_{p^*(i)}\right)+\left(X_i\otimes\bar{X}_{p(i)}\right)'\zeta\end{aligned} \tag{21}$$

and A_i to equal

$$A_i\overset{\text{def}}{\equiv}\left(V_i-g_1\left(X_i,\bar{X}_{p^*(i)}\right)\right)+\rho\left(\bar{V}_{p(i)}-g_2^*\left(X_i,\bar{X}_{p^*(i)}\right)\right)+U_i. \tag{22}$$

Equations (7), (21), and (22) yield an outcome equation of

$$\begin{aligned}Y_i=&W_i'\delta+(X_i\otimes W_i)'\eta+\left(W_i\otimes\bar{X}_{p(i)}\right)'\lambda\\&+h\left(X_i,\bar{X}_{p^*(i)},\bar{X}_{p(i)}\right)+A_i.\end{aligned} \tag{23}$$

Condition (20) implies that A_i is conditionally mean zero given X_i, $\bar{X}_{p^*(i)}$ and W_i^*.

Summarizing, the experimentally-induced restrictions (10), (11), and (12), and our Assumptions 1 and 2 together imply that

$$\mathbb{E}\left[\,A_i\middle|\, X_i,\bar{X}_{p^*(i)},W_i^*\right]=0. \tag{24}$$

The estimation simplifies if we impose a restriction on the peer attrition/noncompliance that is similar to Assumption 2, but is on the observable peer average:

$$\mathbb{E}\left[\bar{X}_{p(i)}-\bar{X}_{p^*(i)}\middle|\, X_i,\bar{X}_{p^*(i)},W_i^*\right]=\mathbb{E}\left[\bar{X}_{p(i)}-\bar{X}_{p^*(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]. \tag{25}$$

This restriction is directly testable. By Equation (25),

$$\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)},W_i^*\right]=\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right].$$

If restriction (25) holds the outcome equation is as in (23), but with a redefined nonparametric h that is a function of X_i and $\bar{X}_{p^*(i)}$ only:

$$\begin{aligned}h\left(X_i,\bar{X}_{p^*(i)}\right)\overset{\text{def}}{\equiv}&\alpha+X_i'\beta+g_1\left(X_i,\bar{X}_{p^*(i)}\right)+\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]'\gamma\\&+\rho g_2^*\left(X_i,\bar{X}_{p^*(i)}\right)+\left(X_i\otimes\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]\right)'\zeta,\end{aligned} \tag{26}$$

and A_i is

$$\begin{aligned}A_i\overset{\text{def}}{\equiv}&\left(V_i-g_1\left(X_i,\bar{X}_{p^*(i)}\right)\right)+\rho\left(\bar{V}_{p(i)}-g_2^*\left(X_i,\bar{X}_{p^*(i)}\right)\right)\\&+U_i+\left(\bar{X}_{p(i)}-\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]\right)'\gamma\\&+\left(X_i\otimes\left(\bar{X}_{p(i)}-\mathbb{E}\left[\,\bar{X}_{p(i)}\middle|\, X_i,\bar{X}_{p^*(i)}\right]\right)\right)'\zeta.\end{aligned} \tag{27}$$

The conditional moment restriction in Equation (24) also holds for this error.

Equations (23) and (24) jointly define a partially linear model with an endogenous parametric component. This is a well-studied semiparametric model (see, e.g., Chen, Linton, and Van Keilegom 2003). The parameters δ, η, and λ are identified; $h\left(X_i,\bar{X}_{p^*(i)}\right)$ is a nonparametric nuisance function.

We implement the partial linear IV estimator using the following approximation for $h(x,\bar{x})$:

$$h\left(X_i,\bar{X}_{p^*(i)}\right)\approx X_i'b+\bar{X}_{p^*(i)}'d+\left(X_i\otimes\bar{X}_{p^*(i)}\right)'f.$$

For this approximation we estimate δ, η, and λ by linear IV fit of Y_i onto a constant, X_i, $\bar{X}_{p^*(i)}$, $\left(X_i\otimes\bar{X}_{p^*(i)}\right)$, W_i, $(X_i\otimes W_i)$, and $\left(W_i\otimes\bar{X}_{p(i)}\right)$ using the excluded instruments W_i^*, $\left(X_i\otimes W_i^*\right)$, and $\left(W_i^*\otimes\bar{X}_{p^*(i)}\right)$. Note that both assigned and realized peer groups enter the main equation.

As in the case with perfect compliance, we do not identify β, γ, and ζ, reflecting the inherent inability of the MET experiment to inform peer group effects. Nevertheless knowledge of δ, η, and λ is sufficient to identify the class of reallocation effects we focus upon.

3. Empirical Implementation and Results

3.1. Data and Sample

The MET study was conducted during the 2009/2010 and 2010/2011 school years in elementary, middle, and high schools located in six large urban school districts in the United States. Its goal was to examine determinants and consequences of teacher quality and teaching practices. In the first study year, MET researchers collected baseline information on teaching practices for each teacher and baseline performance measures for each student. Prior to the second study year, the teachers were randomly assigned to pre-composed classrooms within school-by-grade-by-subject cells (henceforth "randomization blocks"). In many of the elementary schools, the randomization took place within school-by-grade cells in practice to ensure that a given classroom would be taught by the same teacher in both math and ELA. Randomization blocks typically consisted of 2–3 classrooms each, and classroom composition was not manipulated as part of the study. For details about the study design, see White et al. (2019) and Kane et al. (2013).

As a measure of teaching practices, we use Danielson's (2011) "Framework for Teaching" (FFT) measure, collected at baseline (school year 2009/2010). The FFT seeks to capture "what teachers should know and be able to do in the exercise of their profession" (Danielson 2011). Its focus lies on the teachers' ability to actively engage the students in the learning process. The FFT is an observational measure: teachers' classroom interactions were video-taped at different points in time during the school year, four times on average in the baseline year. Subsequently, these videos were rated—in eight different rubrics—by trained raters on a four-point scale (unsatisfactory, basic, proficient, distinguished). We aggregate the ratings across rubrics, videos, and raters. The FFT is positively correlated with a range of teacher quality measures (see Table C.3 in the supplementary materials). We also conducted robustness checks using alternative measures of teacher quality. The results are similar across the different measures (see Section B.2.2 in the supplementary materials).

As outcome measures we use students' 2010/2011 test scores in math and ELA. The scores are based on students' rank in end-of-year state tests and provided in the data as z-scores, that is, they have a mean of zero and a unit standard deviation within each district, subject, and grade.

The dataset also includes background characteristics from school district records for students (age, gender, race/ethnicity, special-education status, free/reduced-price lunch eligibility, gifted status, and whether a student is an English language learner) and for teachers (education, teaching experience in the district). Through section identifiers, we can match students to their classroom peers.

In constructing the estimation sample, we closely follow Garrett and Steinberg (2015). We restrict the sample to all elementary- and middle-school students (grades 4–8) who took part in the randomization. The final sample consists of about 8500 students and 614 teachers in math and of about 9600 students and 649 teachers in ELA (see Section C in the supplementary materials for details).

We discretize both teacher FFT and students' baseline test scores. The FFT is on average 2.5 in math and 2.6 in ELA, which corresponds to a rating between "basic" and "proficient." We set cutoffs at 2.25 and 2.75 (see Figure A.1 in the supplementary materials); using this categorization, 65% of randomization blocks in math and 70% of randomization blocks in ELA include teachers with different levels of FFT (e.g., both a low- and a middle-FFT teacher). The results are not sensitive to the exact position of the cutoffs. Similarly, we split students' 2009/2010 baseline test scores into three bins, corresponding to terciles of the z-score distribution. To include classroom peers into the analysis, we compute the fraction of each student's classmates with high, middle, and low baseline test scores (leave-own-out means). Some teachers and students switched classrooms or schools after the researchers had determined the random assignment. This was partly due to planning uncertainty: the random assignment took place several weeks before the start of the school year, when the classroom and teacher rosters were still subject to change (see Section C.1 in the supplementary materials for details). In the sample, 69% of the students in math and 73% of the students in ELA were actually taught by their randomly assigned teachers.

3.2. Tests of Identifying Assumptions and Restrictions

To test whether the randomization was successful in balancing student characteristics across teachers with different levels of FFT we regress the FFT of a student's assigned teacher on the student's own characteristics, controlling for randomization block fixed effects. None of the student characteristics predict assigned teachers' FFT, individually or jointly, which confirms that the randomization indeed worked (see Table A.1 in the supplementary materials).

Covariate balance, however, is not a sufficient condition to identify reallocation effects under noncompliance by both students and teachers (see Section 2). Assumption 1 requires that own and assigned peer and teacher observables should not predict the difference between realized and assigned unobserved teacher quality. We assess the plausibility of this assumption by using those teacher background characteristics that are not part of the model as replacements for the unobserved quality of a teacher: a teacher's demographics, experience, and education. We regress the difference between realized and assigned teacher characteristics on the student's baseline test score, the FFT of the assigned teacher, and the average baseline test score of the assigned peers. Consistent with Assumption 1, these variables do not jointly predict differences between the characteristics of the assigned and realized teacher in any of the regression fits (see Table A.2 in the supplementary materials).

Similarly, we assess the plausibility of Assumption 2. The assumption states that differences between realized and assigned unobserved peer quality cannot be predicted by assigned teacher observables, conditional on own baseline achievement and the baseline achievement of assigned peers. We regress differences between the assigned and realized characteristics of classroom peers onto the FFT of the assigned teacher, controlling for own baseline test scores and assigned peers' average baseline test scores. Consistent with Assumption 2, teacher FFT does not predict differences between the characteristics of the assigned and realized peers (see Tables A.3 and A.4 in the supplementary materials).

Finally, we test restriction (25), which implies that the baseline test scores of realized peers cannot be predicted by assigned teacher FFT. We regress the baseline test scores of realized peers onto a student's baseline test score, the baseline test scores of her assigned peers, and the FFT of her assigned teacher. We find that this restriction also holds in the data (see Table A.5 in the supplementary materials).

3.3. Computation

3.3.1. Computation of the Optimal Allocation

What is meant by an "optimal" assignment depends on the objective function. We choose to maximize aggregate outcomes (i.e., the sum of all students' test scores). This is a natural point of departure since policy analyses often start with the computation of average effects. Moreover, it is straightforward to compute, justifiable from a utilitarian perspective, and easy to interpret. Our analysis can be modified to accommodate other objective functions. In practice, for example, school principals may care about both the aggregate outcome as well as inequality. Our intention is not to advocate for maximization of the aggregate outcome in practice; rather this makes our analysis comparable to that of other educational policy evaluations. We begin by estimating the parameters of the educational production function based on Equation (23), separately for math and ELA. Since randomization was carried out within randomization blocks, we additionally include randomization block fixed effects in this regression model. We estimate the model's parameters by the method of instrumental variables (IV). We then use the *estimated* parameters, specifically $\hat{\delta}$, $\hat{\eta}$, and $\hat{\lambda}$, to compute three counterfactual outcomes for each student i, that is, her predicted outcome when taught by a low-, middle-, or high-FFT teacher, leaving the original classroom composition unchanged. Aggregated to the classroom level this yields three counterfactual classroom-level test score aggregates. By aggregating the counterfactual outcomes to the *classroom* level, we transform the assignment problem from a many-to-one matching problem to a one-to-one matching problem. This approach is suitable because the configuration of students across classrooms remains fixed; we only consider the effects of reassigning teachers across existing classrooms of students. The one-to-one matching problem is a special linear program, a transportation problem, which is easily solvable.

We impose a few additional constraints to make the reallocation exercise realistic. First, we do not allow teachers to be reassigned across districts or across school types (e.g., to move from an elementary to a middle school). We also present, as a sensitivity check, a version of the allocation where we only allow teachers to be reassigned within their randomization block (see Section B.2 in the supplementary materials). Second, there are a few teachers that teach several sections of a class. In this case, we treat these sections as clusters and allocate one teacher to all sections in each such cluster.

In addition to the optimal assignment, we also compute a "worst" possible assignment, that is, an assignment that minimizes aggregate test scores. The difference between the aggregate outcome for the best and worst assignment is the maximal reallocation gain. This provides an upper bound on the magnitude of student achievement gains that teacher reassignments based on FFT and prior student achievement might yield in practice.

3.3.2. Computation of Average Reallocation Effects

An individual reallocation effect, or reallocation gain, is defined as the difference between an individual student's outcome under two assignments. For example, one can compute, for each student, $\widehat{Y}_i(\tilde{W}_i^{\text{opt}})$, the predicted outcome under the optimal allocation, where $\tilde{W}_i^{\text{opt}}$ takes the values w_L, w_M, or w_H, depending on whether the student would be assigned to a low-, middle-, or high-FFT teacher in an optimal allocation. Similarly, one can compute the same parameter for each student under the status quo, which we denote as $\widehat{Y}_i(W_i)$. An individual reallocation gain can then be computed as the difference between these two outcomes, $\widehat{Y}_i(\tilde{W}_i^{\text{opt}}) - \widehat{Y}_i(W_i)$.

These individual gains can be aggregated in many ways to create policy-relevant parameters. We define our key parameter of interest, the *average reallocation effect*, as

$$\widehat{\text{ARE}} = \frac{1}{N}\sum_{i=1}^{N}\left[\widehat{Y}_i(\tilde{W}_i^{\text{opt}}) - \widehat{Y}_i(W_i)\right]. \tag{28}$$

We also consider three other parameters, (i) conditional average reallocation effects, that is, average reallocation effects for students with varying baseline characteristic x (here, students with low, middle, and high baseline test scores), (ii) the reallocation effect *conditional on being reassigned*, and (iii) the reallocation effect when comparing the optimal to the worst allocation. See Section D.3 in the supplementary materials for a formal definition of these parameters. Notice that, when computing the reallocation effects, we abstract from noncompliance to the optimal allocation. That is, we compare the status quo outcome to the outcome of the optimal allocation under the assumption that there is full compliance to the optimal allocation. It is difficult to predict the degree and nature of noncompliance patterns in the counterfactual scenario, because the experiment we analyze is not a reallocation experiment. The reallocation effects might be smaller in practice due to noncompliance of teachers or students.

3.3.3. Inference

We use the Bayesian bootstrap to quantify our (posterior) uncertainty about AREs. We treat each teacher-classroom pair as an iid draw from some unknown (population) distribution. Following Chamberlain and Imbens (2003), we approximate this unknown population by a multinomial, to which we assign an improper Dirichlet prior. This leads to a posterior distribution which (i) is also Dirichlet and (ii) conveniently only places probability mass on data points observed in our sample. Since our approach to inference is Bayesian the "standard errors" we present for our ARE estimates summarize dispersion in the relevant posterior distribution, not variability across repeated samples. Given that our primary exercise involves solving a social planning problem, the Bayesian approach is both principled and convenient.

We emphasize that our measures of parameter uncertainty have unknown frequentist properties. Consider a scenario in

Table 2. IV regression results of the 3 × 3 model.

		(1)	(2)	(3)	(4)	(5)	(6)
		A. Only teacher effects		B. Full model		C. Without teacher × peer interactions	
		Math	ELA	Math	ELA	Math	ELA
δ	FFT middle	0.069	−0.029	−0.145	−0.219	0.027	−0.082
		(0.053)	(0.048)	(0.138)	(0.173)	(0.060)	(0.059)
	FFT high	0.038	−0.058	−0.547	−0.137	−0.155	−0.154*
		(0.067)	(0.065)	(0.380)	(0.203)	(0.103)	(0.083)
η	FFT middle						
	× baseline middle			0.040	0.057	0.052	0.067
				(0.060)	(0.067)	(0.060)	(0.063)
	× baseline high			0.015	0.076	0.050	0.097
				(0.084)	(0.081)	(0.076)	(0.082)
	FFT high						
	× baseline middle			0.184**	0.150**	0.196**	0.127*
				(0.082)	(0.074)	(0.084)	(0.072)
	× baseline high			0.226**	0.187**	0.265**	0.149
				(0.099)	(0.092)	(0.100)	(0.095)
λ	FFT middle						
	× fraction peers middle			0.318	0.288		
				(0.299)	(0.401)		
	× fraction peers high			0.239	0.161		
				(0.205)	(0.288)		
	FFT high						
	× fraction peers middle			0.641	0.227		
				(0.513)	(0.389)		
	× fraction peers high			0.460	−0.355		
				(0.409)	(0.343)		
β	Baseline middle	0.888***	0.739***	0.843***	0.699***	0.840***	0.682***
		(0.077)	(0.084)	(0.092)	(0.096)	(0.092)	(0.094)
	Baseline high	1.622***	1.555***	1.578***	1.503***	1.550***	1.470***
		(0.103)	(0.105)	(0.124)	(0.118)	(0.119)	(0.115)
R^2		0.617	0.554	0.616	0.552	0.617	0.555
N		8534	9641	8534	9641	8534	9641

Note: The dependent variables are subject-specific test score outcomes in math and ELA. The instrumental variables are based on assigned teacher FFT (Panels A–C) and assigned peer baseline test scores (Panel B). All regressions control for the $h(x, \bar{x})$ function (see Section 2) and for randomization block fixed effects (239 randomization blocks in math and 252 randomization blocks in ELA). Analytic standard errors, clustered by randomization block, are in parentheses.
***significant at the 1%-level **significant at the 5%-level *significant at the 10%-level.

which there are no actual reallocation effects, as when, for example, the educational production function is additively separable in student and teacher attributes. In this case the average outcome at the *estimated* optimal assignment is, essentially, the maximum of a vector of mean zero random variables. This maximum will, in small samples, be biased above the true effect of zero. The frequentist coverage properties of Bayesian credibility sets will also likely be poor in such a setting (see Graham, Imbens, and Ridder 2007; Graham 2011; Andrews, Kitagawa, and McCloskey 2021, for related examples, discussion and results). Developing tractable inference methods for the value function of linear programs is an ongoing area of research (e.g., Hsieh, Shi, and Shum 2022).

3.4. Results

3.4.1. Regression Results

In Table 2 we report IV estimates of our preferred model with three levels of teacher FFT and three levels of student baseline achievement. This corresponds to Equation (23) with both W_i and X_i consisting of dummy variables for middle and high FFT and baseline test scores, respectively. We estimate the model separately for math and ELA test scores and include school-by-grade fixed effects. Section D.1 in the supplementary materials displays the full regression specification.

In this specification, the teacher FFT main effects are insignificant. We do, however, find complementarity between teacher FFT and student baseline scores. These are significant for the high-FFT teachers (i.e., teachers with a "proficient" score on average). Students with middle or high baseline test score levels score significantly higher on end-of-year achievement tests when matched with a high-FFT teacher, compared to students with low baseline test scores (see Table 2, columns 3–6).

To compute optimal assignments and average reallocation effects, we use the IV estimates presented in columns 5–6 of Table 2. These specifications omit the FFT-by-peer composition interaction terms, whose coefficients are poorly determined in all specifications (whether fitted by IV or OLS). Omitting these terms has little effect on either the location or the precision of the coefficients on the FFT-by-baseline interactions. In Tables A.8 and A.9 in the supplementary materials we also present AREs based upon the specifications in columns 3–4 of Table 2. These ARE estimates are larger, albeit less precisely determined. As an additional check, we also compute ITT-based AREs, which are smaller but more precisely estimated than the AREs based on the IV estimates (see Section B.2.4 in the supplementary materials).

3.4.2. Average Reallocation Effects

Our primary reassignment policy considers reassignments of teachers across classrooms within districts and schooling levels

Table 3. Average reallocation gains in math.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A. Optimal versus status quo							
	A.I Full sample				A.II Conditional on being reallocated			
	All students	High	Middle	Low	All students	High	Middle	Low
Gain	0.017	0.028	0.012	0.014	0.036	0.059	0.028	0.026
SE	(0.006)	(0.014)	(0.008)	(0.011)	(0.012)	(0.029)	(0.019)	(0.019)
N	8534	2332	3108	3094	4107	1121	1380	1606
	Panel B. Optimal versus worst allocation							
	B.I Full sample				B.II Conditional on being reallocated			
	All students	High	Middle	Low	All students	High	Middle	Low
Gain	0.040	0.072	0.019	0.038	0.060	0.103	0.032	0.053
SE	(0.012)	(0.034)	(0.015)	(0.026)	(0.019)	(0.048)	(0.022)	(0.037)
N	8534	2332	3108	3094	5746	1626	1912	2208

Note: The table shows average reallocation gains (optimal versus random assignment in Panel A and optimal versus worst assignment in Panel B). The gains are expressed in test score standard deviations. The computations are based on a 3×3 model without teacher-by-peer interactions (see Table 2, column 5). The results are presented separately for the full sample of students (columns 1–4) and for the sample of classrooms that would get a new teacher as a result of the reallocation (columns 5–8). The reassignments are carried out within school types and districts. Standard errors are in parentheses and computed using the Bayesian bootstrap with 1000 replications (see Section 3.3.3). High/middle/low: students in the top/middle/bottom tercile of the baseline test score distribution.

(elementary and middle schools). Under this scenario we find that, when moving from the status quo to an optimal assignment, 49% of the students in the math sample and 47% of the students in the ELA sample are assigned to a new teacher.

In the math sample (Table 3), the optimal allocation improves average test scores by 1.7% of a test score standard deviation compared to the status quo. This effect is precisely determined with a Bayesian bootstrap standard error of 0.6%. The reported effects are largely driven by students with high baseline test scores. These students gain 2.8% of a test score standard deviation on average; students with middle and low baseline test scores, in contrast, gain 1.2% and 1.4% of a test score standard deviation, respectively (on average). Since only half of the students experience a change in their teacher, the average effect represents an equal-weighted mixture of a zero effect and a positive effect on those students who do experience a change in teachers. The average effect for the latter group is 3.6% of a test score standard deviation (Panel A.II, SE = 1.2%).

The comparison of an optimal allocation with a worst allocation yields improvements that are about twice as large. Relative to a worst allocation, an optimal allocation improves test scores by 4.0% of a standard deviation on average (SE = 1.2%). The gains are 7.2% of a standard deviation for students with high baseline test scores and 1.9% and 3.8% of a standard deviation for students with middle and low baseline test scores, respectively. If one considers only those students who are reassigned to a new teacher, the reallocation effect amounts to 6.0% of a standard deviation on average (Panel B.II, SE = 1.9%).

One way to benchmark the magnitude of these AREs is to compare them with the effects of hypothetical policies aimed at improving teacher value-added measures (VAMs). Such policies are controversial, as is the evidence marshaled to support them. Here we offer no commentary on the advisability of actually adopting VAM-guided teacher personnel policies; nor do we offer an assessment of VAM studies. Rather we simply use these studies, and the policy thought experiments they motivate, to benchmark the ARE findings.

Teacher value-added is typically conceptualized as an invariant intercept-shifter, which uniformly raises or lowers the achievement of all students in a classroom. In this framework replacing a low value-added teacher with a high one will raise achievement for all students in a classroom. Rockoff (2004) estimates that the standard deviation of the population distribution of teacher value-added (in a New Jersey school district) is around 0.10 test score standard deviations in both math and reading. Recent studies find somewhat higher estimates: Chetty, Friedman, and Rockoff (2014) estimates that the standard deviation of teacher value-added is 0.16 in math and 0.12 in reading; similarly, Rothstein (2017) finds values of 0.19 in math and 0.12 in reading.

Using a standard deviation of 0.15 we can consider the effect of a policy which removes the bottom $\tau \times 100\%$ of teachers, sorted by VAM, and replaces them with teachers at the $\tilde{\tau}$th quantile of the VAM distribution. Under normality the effect of such a policy on average student achievement is to increase test scores by

$$(1-\tau)\,\sigma\,\frac{\phi\left(\frac{q_\tau}{\sigma}\right)}{1-\Phi\left(\frac{q_\tau}{\sigma}\right)}+\sigma\,\Phi^{-1}\left(\tilde{\tau}\right)$$

standard deviations. Setting $\tau = 0.05$ and $\tilde{\tau} = 0.75$ this expression gives an estimate of the policy effect of 0.021 (i.e., 2.1% of a test score standard deviation). This is comparable to the average effect on math achievement associated with moving from the status quo MET assignment to an optimal one. In practice correctly identifying, and removing from classrooms, the bottom 5% of teachers would be difficult to do. Replacing them with teachers in the top quartile of the VAM distribution even more so (see, e.g., Staiger and Rockoff 2010; Jackson, Rockoff, and Staiger 2014, for discussions of VAM-guided policies). Contextualized in this way the AREs we find are large.

An attractive feature of the policies we consider is that they are based on measurable student and teacher attributes, not noisily measured latent ones. At the same time we are mindful that most school districts would not find it costless to reallocate teachers freely across classrooms and schools (see Glazerman et al. 2013). Moreover, observational measures such as the FFT can also contain measurement error. Further research is needed to determine both the costs of reassignment policies and the set of variables that such policies should be based upon.

For ELA achievement we find smaller reallocation effects (see Table A.10 in the supplementary materials). Moving from

the status quo to an optimal allocation is estimated to raise achievement by 0.8% of a test score standard deviation (SE = 0.6%). As with math, these gains are concentrated among students with high baseline scores who are assigned a new teacher. These students experience an average gain of 5% of a test score standard deviation (SE = 2.5%).

In sum, the optimal assignment we consider yields hypothetical improvements in test score outcomes across the distribution of student baseline achievement. Yet, while the gains are large for students with high baseline test scores—up to 10% of a standard deviation for students with high baseline test scores in math—for students with middle and low baseline test scores, the gains are smaller. Consequently, the reallocation does not narrow achievement gaps between students with high baseline test scores and those with middle or low baseline test scores. Moreover, reallocations seem to matter more in math, compared to ELA. This is in line with the value-added literature, which consistently reports higher value-added in math compared to ELA.

4. Conclusion

We provide an econometric framework that allows us to semiparametrically characterize complementarity between teaching practices and student baseline test scores. Our framework exploits the random assignment of teachers to classrooms available in the MET dataset while formally dealing with noncompliance by both teachers and students.

The results provide strong evidence of complementarity between student baseline test scores and teacher practices. This complementarity, if taken into account when assigning teachers to classrooms, appears to make a difference in students' performance across the distribution of baseline achievement. Teacher reassignments that maximize the average test score, however, will not close the achievement gap between students with low versus high baseline achievement levels. They actually widen this achievement gap. A different objective function, one that values equity as well as efficiency, may be preferred in practice. It is also possible that assignments based on other student and teacher attributes (e.g., race or gender) might both raise average achievement and narrow achievement gaps.

Supplementary Materials

The authors provide the following supplementary materials:

1. An online appendix containing supplementary results, information on the data and sample characteristics, and details about the empirical procedures.
2. A replication package containing descriptions and code to reproduce Tables 2 and 3 from the original data.

Acknowledgments

We thank audiences at the 2017 All-California Econometrics Conference at Stanford University, the 2019 Bristol Workshop on Economic Policy Interventions and Behavior, the 2020 IZA Workshop on the Economics of Education, the 2021 FAIR workshop at NHH Norway, SOLE 2021, IAAE 2021, EEA 2021, VfS 2021, EALE 2021, UC Riverside, NESG, University of Duisburg-Essen, Tinbergen Institute Amsterdam, Lund University, IFN Stockholm, IFAU Uppsala, and the University of Southern California for helpful feedback. We thank Tommy Andersson, Kirabo Jackson, Magne Mogstad, Hessel Oosterbeek, Daniele Paserman, and Hashem Pesaran for useful comments and discussions. All the usual disclaimers apply. Thiemann was affiliated with USC Dornsife INET while working on this article.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

Financial support for Graham was provided by the National Science Foundation (SES #1357499, SES #1851647).

ORCID

Bryan S. Graham http://orcid.org/0000-0002-9637-5282
Geert Ridder http://orcid.org/0000-0003-4828-1617
Petra Thiemann http://orcid.org/0000-0003-2295-8706
Gema Zamarro http://orcid.org/0000-0002-3893-7523

References

Ai, C., and X. Chen (2003), "Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions," *Econometrica*, 71, 1795–1843. [1329]

Andrews, I., Kitagawa, T., and McCloskey, A (2021), "Inference on Winners," manuscript. [1337]

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., and Schady, N. (2016), "Teacher Quality and Learning Outcomes in Kindergarten," *The Quarterly Journal of Economics*, 131, 1415–1453. [1329]

Aucejo, E., Coate, P., Fruehwirth, J. C., Kelly, S., and Mozenter, Z. (2022), "Teacher Effectiveness and Classroom Composition," *The Economic Journal*. *https://doi.org/10.1093/ej/ueac046*. [1329]

Bhattacharya, D. (2009), "Inferring Optimal Peer Assignment from Experimental Data," *Journal of the American Statistical Association*, 104, 486–500. [1329,1332]

Burgess, S., Rawal, S., and Taylor, E. S. (2021), "Teacher Peer Observation and Student Test Scores: Evidence from a Field Experiment in English Secondary Schools," *Journal of Labor Economics*, 39, 1155–1186. [1329]

Carrell, S. E., Sacerdote, B. I., and West, J. E. (2013), "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81, 855–882. [1329]

Chamberlain, G., and Imbens, G. W. (2003), "Nonparametric Applications of Bayesian Inference," *Journal of Business and Economic Statistics*, 21, 12–18. [1336]

Chen, X., Linton, O., and Van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608. [1334]

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2012), "Great Teaching," *Education Next*, 12, 59–64. [1328]

——— (2014), "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, 104, 2593–2632. [1332,1338]

Cohen-Vogel, L., Feng, L., and Osborne-Lampkin, L. (2013), "Seniority Provisions in Collective Bargaining Agreements and the 'teacher quality gap'," *Educational Evaluation and Policy Analysis*, 35, 324–343. [1329]

Danielson, C. (2011), "The Framework for Teaching Evaluation Instrument, 2011 Edition," Retrieved on May 31, 2020, from *https://danielsongroup.org/downloads/2011-framework-teaching-evaluation-instrument*. [1328,1335]

Darling-Hammond, L. (2015), "Can Value Added Add Value to Teacher Evaluation?" *Educational Researcher*, 44, 132–137. [1328]

Dee, T. S. (2004), "Teachers, Race, and Student Achievement in a Randomized Experiment," *Review of Economics and Statistics*, 86, 195–210. [1328]

Garrett, R., and Steinberg, M. P. (2015), "Examining Teacher Effectiveness Using Classroom Observation Scores: Evidence From the Randomiza-

tion of Teachers to Students," *Educational Evaluation and Policy Analysis*, 37, 224–242. [1329,1335]

Glazerman, S., Protik, A., Teh, B.-r., Bruch, J., and Max, J. (2013), "Transfer Incentives for High-Performing Teachers: Final Results from a Multisite Randomized Experiment. NCEE 2014-4003." *National Center for Education Evaluation and Regional Assistance*. [1338]

Graham, B. S. (2008), "Identifying Social Interactions through Conditional Variance Restrictions," *Econometrica*, 76, 643–660. [1330,1332]

——— (2011), "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," in *Handbook of Social Economics* (Vol. 1B), eds. J. Benhabib, A. Bisin, and M. O. Jackson, pp. 965–1052, Amsterdam: North-Holland. [1337]

Graham, B. S., Imbens, G. W., and Ridder, G. (2007), "Redistributive Effects for Discretely-Valued Inputs," IEPR Working Paper 07.7, University of Southern California. [1328,1330,1332,1337]

——— (2010), "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," Working Paper 16499, NBER. [1330]

——— (2014), "Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis," *Quantitative Economics*, 5, 29–66. [1328,1329,1330,1331]

——— (2020), "Identification and Efficiency Bounds for the Average Match Function Under Conditionally Exogenous Matching," *Journal of Business & Economic Statistics*, 38, 303–316. [1328,1329,1330]

Grissom, J. A., Kalogrides, D., and Loeb, S. (2015), "The Micropolitics of Educational Inequality: The Case of Teacher–Student Assignments," *Peabody Journal of Education*, 90, 601–614. [1329]

Hanushek, E. (1971), "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data," *American Economic Review*, 61, 280–288. [1328]

Hanushek, E. A., Kain, J. F., and Rivkin, S. G. (2004), "Disruption versus Tiebout Improvement: The Costs and Benefits of Switching Schools," *Journal of Public Economics*, 88, 1721–1746. [1332]

Hsieh, Y.-W., Shi, X., and Shum, M. (2022), "Inference on Estimators Defined by Mathematical Programming," *Journal of Econometrics*, 226, 248–268. [1337]

Jackson, C. K., Rockoff, J. E., and Staiger, D. O. (2014), "Teacher Effects and Teacher-Related Policies," *Annual Review of Economics*, 6, 801–825. [1338]

Kalogrides, D., Loeb, S., and Beteille, T. (2011), "Power Play? Teacher Characteristics and Class Assignments," Working Paper 59, National Center for Analysis of Longitudinal Data in Education Research. [1329]

Kane, T. J., McCaffrey, D. F., Miller, T., and Staiger, D. O. (2013), "Have we Identified Effective Teachers? Validating Measures of Effect Teaching Using Random Assignment," Met Project Research Paper, Bill & Melinda Gates Foundation. [1335]

Loeb, S., Soland, J., and Fox, L. (2014), "Is a Good Teacher a Good Teacher for All? Comparing Value-Added of Teachers with Their English Learners and Non-English Learners," *Educational Evaluation and Policy Analysis*, 36, 457–475. [1328]

Manski, C. F. (1993), "Identification of Endogenous Social Effects: The Reflection Problem," *Review of Economic Studies*, 60, 531–542. [1332]

McFarland, J., Hussar, B., Zhang, J., Wang, X., Wang, K., Hein, S., Diliberti, M., Forrest Cataldi, E., Bullock Mann, F., and Barmer, A. (2019), "The Condition of Education 2019," NCES 2019-144, National Center for Education Statistics. *https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2019144* [1328]

Morganstein, D., and Wasserstein, R. (2014), "ASA Statement on Value-Added Models," *Statistics and Public Policy*, 1, 108–110. [1328]

Robinson, P. M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954. [1333]

Rockoff, J. E. (2004), "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94, 247–252. [1329,1338]

Rothstein, J. (2010), "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement 2010," *Quarterly Journal of Economics*, 125, 175–214. [1329]

——— (2017), "Measuring the Impacts of Teachers: Comment," *American Economic Review*, 107, 1656–84. [1338]

Snyder, T. D., de Brey, C., and Dillow, S. A. (2017), "Digest of Education Statistics," NCES 2018-070, National Center for Education Statistics. *https://nces.ed.gov/pubs2018/2018070.pdf* [1328]

Staiger, D. O., and Rockoff, J. E. (2010), "Searching for Effective Teachers with Imperfect Information," *The Journal of Economic Perspectives*, 24, 97–117. [1338]

U.S. Department of Education (2020), "Teacher Performance Evaluations in U.S. Public Schools," NCES 2020-133. Available at *https://nces.ed.gov/pubs2020/2020133.pdf*, Accessed January 21, 2021. [1329]

White, M., Rowen, B., Alter, G., Blankenship, L., Greene, C., and Windish, S. (2019), *User Guide to Measures of Effective Teaching Longitudinal Database (MET LDB)*, Ann Arbor: Inter-University Consortium for Political and Social Research, The University of Michigan. [1335]