ISSN 2042-2695

CENTRE *for* ECONOMIC PERFORMANCE

CEP Discussion Paper No 1574

Revised April 2019
(Replaced September 2018 version)

Teacher Effectiveness and Classroom Composition

Esteban M. Aucejo. Patrick Coate
Jane Cooley Fruehwirth, Sean Kelly
Zachary Mozenter

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract
This paper studies how the effectiveness of teachers varies by classroom composition, combining random assignment of teachers with rich measures of teaching practices based on a popular teacher-evaluation protocol. We find that some teaching practices are more effective in raising math achievement in classrooms with higher average prior achievement, and others are more effective in classrooms with less heterogeneity in prior achievement. We use these estimates to simulate the effects of reallocating classrooms among teachers within schools. We find substantial differences between counterfactual and actual teacher effectiveness rankings, supporting the importance of classroom composition for evaluating teachers and prescribing practice.

Key words: teacher, practices, peer effects, effectiveness
JEL Codes: I2; I20; I21

This paper was produced as part of the Centre's Education and Skills Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170269 to University of North Carolina, Chapel Hill. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. Fruehwirth also thanks the British Academy and the Leverhulme Trust's Philip Leverhulme Prize. We thank seminar participants at the Kentucky-Cleveland Fed Workshop on the Economics of Education, and NBER Summer Institute. We also thank Pat Bayer, Ken Bollen, Cassie Guarino, David Guilkey, Laura Hamilton, Joe Hotz, Ju Hyun Kim, Spyros Konstantopolous, Lindsay Matsumara, Doug Staiger, Greg Veramendi, Valentin Verdier for helpful comments and conversations.

Esteban M. Aucejo, Arizona State University and Centre for Economic Performance, London School of Economics. Patrick Coate, National Council on Compensation Insurance. Jane Cooley Fruehwirth, University of North Carolina. Sean Kelly, University of Pittsburg. Zarhary Mozenter, University of North Carolina.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Teachers are largely understood to be the most important school-level determinant of achievement. While a vast literature has contributed to significant improvements in our understanding of the role of teachers, important gaps in our knowledge remain, including what makes a teacher effective and how to prescribe teaching practices that will meet the needs of all students in the classroom (Gamoran et al., 2000; Rivkin et al., 2005). We hypothesize that some of the key challenges in measuring teacher effectiveness and prescribing teaching practice arise due to heterogeneity in the effects of teachers by classroom composition. For instance, the benefits of encouraging classroom discussion may vary depending on the heterogeneity in initial achievement of a student's classmates. Alternatively, teaching practices that create a positive learning environment and promote a learning dialogue among students could amplify the capacity for students to benefit from their peers.[1] Yet, little is known about how or if teacher effectiveness varies by classroom composition.

We address this gap by studying complementarities between teachers and classroom composition. The existence of complementarities has important implications for key teacher-related policies, including the importance of taking classroom composition into account when measuring teacher effectiveness and prescribing teaching practice. It further speaks to continued debates about whether current value-added models of teacher effectiveness can fairly be used in high-stakes settings or whether they unfairly penalize or reward teachers based on the background of students in their classrooms.

We exploit a unique data set to address this question–the Measures of Effective Teaching (MET) Longitudinal Database–which includes rich information on teaching practices in a context where teachers are randomly assigned to classrooms. Teachers are evaluated by trained raters using a research-based protocol that is increasingly used to measure teaching effectiveness in schools nationwide, the *Framework for Teaching Evaluation Instrument (FFT)*

[1]Teaching practices not only involve the principles and methods used for instruction (e.g. class discussions vs. recitation), but also those actions that affect the social dynamics of a given classroom (e.g. classroom management).

(Danielson, 2011).[2] For our measure of classroom composition, we focus on classroom peer initial achievement, the most-studied type of peer spillover in the literature (Sacerdote, 2011).

The random assignment of teachers eliminates one of the most important confounding factors for measuring teacher effectiveness, the systematic matching of students to teachers that would lead us to confound teachers or peer effects with unobservable teacher or peer quality. However, even with random assignment, our identification strategy needs to address a number of remaining endogeneity concerns. The first is that there is considerable non-compliance in the data. We address this by relying on the variation generated by the randomly-assigned teacher rather than the actual teacher. Second, classroom composition may be endogenous. We apply a result in Bun and Harrison (2018) and Nizalova and Murtazashvili (2014) to show that the random assignment of teachers to classrooms is enough to obtain consistent estimates of the complementarities between teaching practice and classroom composition as long as students do not re-sort to classrooms in response to teachers, which we can test using initially-assigned classrooms. Third, if teachers choose practices to maximize student achievement, the observed teaching practice could be endogenous to the classroom composition. We address this primarily by focusing on prior year teaching practices, thus capturing teachers' proclivities toward certain practices. Fourth, teaching practice is measured with error. We exploit multiple measures of teaching practice and show robustness to a number of approaches to dealing with measurement error.

We ground our empirical strategy in a simple theoretical model of student behavior, which helps inform the structure of the estimating equations and illustrates the potential pervasiveness of the complementarities in teaching practice and classroom composition. We show that even when the learning production function does not directly depend on the interaction between teaching practice and peer initial achievement, a complementarity between teachers and peers could emerge indirectly through students' endogenous responses to

[2] Kane et al. (2011) shows the importance of this teacher evaluation protocol in an observational context.

teaching practices. One such example is when teachers with better classroom management practices make engagement less costly, and students benefit more from their peers if they are engaged with the material.

Our main findings indicate that complementarities between classroom composition and teaching practice play a key role in student achievement. We illustrate the importance of complementarities in measuring teacher effectiveness through some simulations. We find that reallocating teachers across classrooms within school has significant effects on measures of teachers' contributions to learning and ranking. The significant fluctuations in measured teacher contributions, even with these relatively small local shifts in classroom composition, points to the need for caution in implementing high-stakes policies based on teacher value-added, such as those that aim to replace the worst 5% to 10% of teachers with average teachers (Chetty et al., 2014; Hanushek, 2011).[3]

To take the additional step beyond informing teacher effectiveness to prescribing teaching practice requires ensuring that our measures of practice are not correlated with other unboservable aspects of the teacher. This is a challenge that all the literature that seeks to evaluate characteristics of effective teaching shares.[4] We find that while certain subdomains of FFT, which we label as classroom management practices, interact positively with average peer initial achievement; other subdomains, which we label as challenge/student-centered practices, interact negatively with the interquartile range of prior achievement. To inform whether our FFT-based measures are confounding other aspects of teacher quality, we show that our results are robust to controlling for an unusually rich set of teacher quality measures, including principal and student surveys along with a teaching knowledge assessment.[5]

We make several important contributions to the literature. First, we demonstrate how failing to capture the heterogeneity in the effectiveness of

[3]Section 7 describes how we define teacher contribution to learning.

[4]For instance, see Araujo et al. (2016) and Taylor (2018) for discussions of this challenge.

[5]Taylor (2018) also shows that different type of instructional methods play an important role on student achievement beyond just teaching skills. Although educational researchers make an important distinction between teacher quality and teaching quality (Hamilton, 2012; Kennedy, 2010), we use the term "teacher" here, assuming the teacher knowledge measures reflect relatively stable traits.

teaching practice by classroom composition leads us to understate the importance of measures of teacher effectiveness and even, in some cases, to infer that the practice does not matter when in fact the effects are sizable in certain classrooms. This provides insight into why observable teacher measures may often do a poor job of capturing teacher quality (e.g. Rivkin et al., 2005). From a policy perspective, understanding this type of heterogeneity is crucial for identifying what teaching practices matter and in what classroom contexts.

Second, our research connects closely to a number of recent studies that consider heterogeneity in teacher effectiveness by student background characteristics (Lavy, 2015; Fox, 2016; Konstantopoulos, 2009).[6] However, by focusing on heterogeneity by classroom composition, our work is substantively different in focus. Furthermore, we show that heterogeneity by classroom composition seems to be of significantly larger magnitudes than heterogeneity by a student's prior achievement.

Third, our study also provides useful complementary evidence to the value-added literature which argues fairly persuasively that teachers matter (Jackson et al., 2014; Koedel et al., 2015; Rivkin et al., 2005; Chetty et al., 2014; Rothstein, 2010). Consistent with our central hypothesis that teacher effectiveness varies with who the teacher teaches, interesting recent work by Stacy et al. (2013) shows that value-added estimates are significantly more stable year-to-year for teachers of students with higher-initial achievement. The most closely related work is an innovative paper by Jackson (2013), which demonstrates a significant role for match quality between teachers and schools. Instead of focusing on value-added measures of teacher effectiveness, we focus on observational rubrics which have several advantages: (1) they capture multi-dimensional aspects of teaching, (2) they move us closer to being able to prescribe practices that matter and (3) they apply to many school districts in the US that are actively using FFT (or FFT-based) protocols to evaluate

[6]For instance, Lavy (2015) finds larger effects of challenge/student-centered teaching for girls and low-SES students. Connor et al. (2004) show larger effects of some types of challenge/student-centered practices for children with higher initial achievement. Finally, Konstantopoulos (2009) finds somewhat larger effects of teacher effectiveness for high-SES students.

teachers.[7]

A number of other studies have used the MET data to generate important insights on the role of observation protocols in measuring effective teaching (Cantrell and Kane, 2013).[8] For instance, Kane et al. (2013) verify that value-added metrics can be effective ways of evaluating teacher effectiveness in observational data and that multiple metrics of teacher effectiveness, including observations of practice, further improve understanding of a teachers' underlying effectiveness. Mihaly et al. (2013) show important commonalities between video observation and value-added measures of effectiveness.

Fourth, our paper builds on the insights of a large peer effects literature by studying how the effects of classroom compostion vary by teaching practice (See Sacerdote, 2011; Epple and Romano, 2010, for reviews). Zimmer (2003) and Duflo et al. (2011) consider heterogeneity in peer effects by student prior achievement and by whether the school tracks or not, which relates to the present study in interesting ways. A handful of recent studies show peer effects may be driven by how the teacher adapts or targets her teaching (Jackson, 2016; Duflo et al., 2011; Lavy et al., 2012; Lee et al., 2014). This paper builds on this literature by considering the moderating role of teaching practice.

The rest of the paper proceeds as follows. We first describe the data in Section 2, including our measures of teaching practice. Section 3 presents our theoretical framework and Section 4 our empirical strategy. Section 5 presents our main findings, followed by an analysis of the possible mechanisms behind our main results in Section 6. Section 7 performs simulation exercises that study how reallocation of teachers into classrooms affects their contribution to learning and their relative rankings. Finally, Section 8 concludes.

[7]While the econometric issues associated with allowing estimated teacher value-added to vary by classroom composition are also of interest, illustrating the potential variation in teacher effectiveness using teacher evaluation protocols is a natural starting point.

[8]Araujo et al. (2016) and Bacher-Hicks et al. (2017), in different settings, also illustrate the importance of teacher observation protocols for measuring teacher effectiveness.

2 Data

The Measures of Effective Teaching (MET) Longitudinal Database provides detailed information on teaching practices, student outcomes, and classroom composition from 5 large urban public school districts in the United States over two academic years (2009-2010 and 2010-2011).[9] The data are linked to district administrative records, which include detailed student information, most importantly, current and prior measures of student achievement, but also age, race/ethnicity, gender, special education status, free lunch eligibility, gifted status, and English language learner status. The data also include rich measures related to teacher aptitude, such as the Content Knowledge for Teaching (CKT) assessment, and school principal evaluations.[10]

We analyze students' math performance because it has traditionally been shown to be more malleable to school inputs. Moreover, we focus on elementary school students (grades four and five) given that most of them are taught by general elementary teachers in self-contained classrooms with more concentrated exposure to the same peers and teachers.[11]

2.1 Measuring Teaching Practice

We make use of a well-known, research-based, classroom-observation protocol that measures teaching practices, the *Framework for Teaching* (FFT). Increasingly school districts have begun to use these types of protocols for teacher evaluation purposes and FFT is the most popular (AIR, 2013). According to

[9]The original 6 districts include New York City Department of Education, Charlotte-Mecklenburg Schools, Denver Public Schools, Memphis City Schools, Dallas Independent School District, and Hillsborough County Public Schools, but only 5 have video observation data of teaching practice. Kane and Staiger (2012) provides a detailed description of how schools were selected to participate in the MET project. More importantly, Kane and Staiger (2012) argues that MET teachers are comparable by most measures to their non-MET peers in the district, suggesting that they are representative of the districts included.

[10]The purpose of the CKT math assessment is to measure knowledge tied to the teaching of mathematics, such as, choosing and using appropriate mathematical representations; choosing examples to illustrate a mathematical concept; interpreting student work, including use of nonstandard strategies; and evaluating student understanding.

[11]Appendix B provides a detailed description of the sample selection and sample characteristics.

MET project (2010), "*FFT has been subjected to several validation studies over the course of its development and refinement, including an initial validation by Educational Testing Service (ETS).*"[12] The protocol divides teaching into four domains and the MET database rates teachers on two of them: *classroom environment* and *instruction*. We observe scores for eight different subdomains of these two domains by a median of seven different highly-trained, independent raters, many of them current or former teachers.[13] These raters had to pass reliability tests in which their scores were compared with master scores on a number of videos. This provides some assurance of the quality of these observational data and help us to address measurement error, as we discuss further in Section 4.

Though FFT was designed so that each subdomain represents a separate aspect of teaching practice, we perform an exploratory factor analysis to determine the number of components that are actually separable in the data. Appendix Table A.2 shows the correlations between the different subdomains and the loadings on each subdomain after performing an oblique rotation of the factors.[14] This analysis suggests that FFT measures can be divided into two separable broad teaching practices. There are five sub-scales which load heavily on the first factor, including *establishing a culture of learning*, *communicating with students*, *engaging students in learning*, *using assessment in instruction* and *using questioning and discussion techniques*. These all reflect what we will call *challenge/student-centered practices* that encourage classroom dialogue and student involvement.[15] The subdomains that load on the

[12] Of the MET observation protocols, two–FFT and CLASS–are generic protocols designed to apply across instruction in a range of subject-matters. In our view, of these, FFT has the most comprehensive architecture capturing teaching practices.

[13] The score assigned to each component ranges between 1 and 4, where each each number refers to a level (1:unsatisfactory, 2:basic, 3:proficient, 4:distinguished). Appendix Table A.1 provides a description of each of the sub-components of the FFT protocol.

[14] The results reported take the average across raters so that there is one observation per component per teacher. Results are similar if we perform the exploratory factor analysis at the level of the rater or if we use orthogonal rotations. They are also similar if we extract rater fixed effects and video quality prior to performing the factor analysis.

[15] We have chosen the term "challenge/student-centered practices" to try to capture the overall emphasis of the model items. These include all of the items in a domain called "Instruction" and one item from the "Classroom environment" domain in the Danielson

second factor are *creating an environment of respect and rapport*, *managing student behaviors* and *managing classroom procedures*. We will refer to these as *classroom management* practices, as they all relate to teaching practices that lead to a better classroom environment. Taken together the factors explain 92% of the total variance in the data.[16]

As a final robustness check, we also implement confirmatory factor analysis with the aim to establish whether the proposed grouping of the FFT subdomains provides a better fit of the data than alternative models. First, we compare our model with a competing specification in which all the FFT subdomains load in only one latent factor. Second, we test our classification with the grouping that has been predetermined in the FFT protocol (i.e., classroom environment and instruction domains).[17] In both cases, the Bayesian information criterion (BIC) indicates that our proposed classification provides a better fit of the data.[18] Our empirical strategy mainly makes use of averages across the sub-scales that according to the exploratory factor analysis correspond to each broad practice (i.e., classroom management and challenge/student-centered practices), but we also explore other ways of addressing measurement error, as described in detail in Section 4.[19]

rubric. We prefer not to use the label *instruction* because it is not descriptive of what the protocol is designating as good instruction. Many of the FFT domains entail elements of student-centered instruction (e.g., in the engaging students in learning domain,"students identify or create their own materials for learning"). Yet, it is important to note that the FFT protocol is well balanced with "challenge" items (e.g. the first indicator of proficiency in the questioning and discussion techniques sub-domain is "questions of high cognitive challenge" (Danielson, 2011).)

[16]An initial exploratory factor analysis shows that there is only one eigenvalue greater than 1, a possible rough rule of thumb for determining the number of factors. However, one factor explains 0.79 of the variation and a second factor explains a substantial additional part, 0.13, which is an additional criteria used to determine the number of factors.

[17]*Classroom environment* includes: environment of respect and rapport, establishing a culture for learning, managing student behaviors, and managing classroom procedures. While *instruction* includes: communicating with students, engaging students in learning, using assessment in instruction, and using questioning and discussion techniques.

[18]This analysis has been performed using the "confa" command in Stata, which deals with problems of identification in factor models (Kolenikov, 2009).

[19]We also replicate our empirical strategy using principal component and following the FFT classification as alternative measures of challenge/student-centered and classroom management practices. Results in all cases are similar.

Table 1: Within and Between-Randomization Block Variation in Classroom Measures

	Mean	Std. Dev.	Min	Max	Std. Dev. Between	Std. Dev. Within
Classroom Composition						
Avg Peer Math$_{t-1}$	0	1	-2.31	3	0.84	0.58
IQR Peer Math$_{t-1}$	0	1	-2.45	2.92	0.78	0.69
Teaching Practices						
Challenge/Student-Centered	0	1	-3.06	2.23	0.74	0.69
Classroom Management	0	1	-3.15	2.25	0.74	0.63

Notes: The sample size is 2632 and focuses on 2010-11 school year when students were randomly assigned within randomization blocks. Teaching practices are measures in $t-1$ based on FFT. The last two columns decompose the standard deviation for each variable into between randomization block and within randomization block components.

2.2 Randomization and Sample Variation

A key aspect of the MET data is that teachers were randomly assigned within school and grade to classrooms of students during the second academic year of the study (2010-2011). This was done within randomization-blocks that were identified by the principal. Appendix A describes this process in more detail. Central to our strategy is to exploit within-randomization block variation in classroom composition and teacher practices. We measure classroom composition in two ways: average prior achievement of current classmates and the interquantile range of prior achievement in the current classroom.[20] Both

[20]We use IQR (i.e. difference in test score performance between the 75th and 25th percentile students in a given class) to measure classroom heterogeneity rather than standard deviation due to the fact that IQR is less sensitive to the presence of outliers, which is a particular concern in a context where classrooms could be small in size. Nevertheless, our main specifications presented in columns (1) and (2) of Table 3 are robust to replacing IQR with the standard deviation.

classroom compostion and teaching practices are normalized within the sample to ease interpretation, as discussed in Section 4. Table 1 documents considerable within-randomization block heterogeneity in both classroom composition and teaching practice.

3 Model

In this section, we motivate how interactions between teaching practice and peer initial achievement arise through a number of intuitive mechanisms. The simplest model has these interactions arising through the production technology. This makes sense for a number of possible teaching practices. For instance, encouraging classroom discussion would create more of a team production climate where peers matter more for each student's achievement. Alternatively, some teaching practices could enter indirectly to the achievement production function through students' behavioral responses (e.g., engagement, attentiveness). In this case, complementarities would arise if good behavior changes whether students benefit from their peers. For instance, classroom management practices could help ensure the necessary behavior to create a good learning environment. While the production technology channel is straightforward, it is helpful to illustrate the behavioral channel with a simple model. The model also informs the empirical specification we take to the data.[21]

Let Y_{it} denote achievement of a student i at time t. Let the index $c_t = c(i,t)$ denote i's classroom in period t and then the vector of classroom peer achievement excluding i is denoted $Y_{-ic_tt} = (Y_{1t}, ..., Y_{i-1,t}, Y_{i+1,t}, ..., Y_{Nt})$. A student's class has a teacher indexed $j = j(i,t)$ who uses teaching practice(s) P_j. We begin with a value-added model where achievement production is a function of prior achievement and some moment of the prior achievement distribution of student i's time t classmates $(m(Y_{-ic_tt-1}))$. We introduce student behav-

[21] We take the teaching practice as given in order to focus on student responses. We can identify most convincingly the effects of a fixed or persistent aspect of teaching practice and postpone considering the endogenous response of teachers to the classroom composition in future work.

ior, b_{it}, conceptualized broadly as behaviors conducive to achievement, such as attentiveness, engagement and/or effort. Achievement production includes direct interactions between teaching practice and classroom composition and the possibility of an indirect channel by allowing the marginal benefits of behavior to vary by the classroom composition, i.e.,

$$\begin{aligned} Y_{it} = \beta_0 + \beta_b b_{it} + \beta_{by} b_{it} Y_{it-1} + \beta_{b\bar{y}} b_{it} m(Y_{-ic_t t-1}) + \beta_y Y_{it-1} + \beta_{\bar{y}} m(Y_{-ic_t t-1}) + \\ + \beta_p P_j' + \beta_{py} P_j' Y_{it-1} + \beta_{p\bar{y}} P_j' m(Y_{-ic_t t-1}) + \epsilon_{it}, \quad (1) \end{aligned}$$

where ϵ_{it} denotes the residual.

Note that when $m(Y_{-ic_t t-1})$ is equal to average peer prior achievement this form of the achievement production function is comparable to the classic achievement production function with peer spillovers that is generally the focus of the literature (Sacerdote, 2011). It is augmented with controls for measures of teaching practice and teaching practice interacted with peers. This model is understood to capture the reduced-form effect of peers, inclusive of both endogenous (effects arising through simultaneity in achievement) and contextual effects (arising through direct spillovers from peer prior achievement), as in the terminology set forth by Manski (1993). We begin here as it is the most straightforward model to connect with the literature, but we illustrate other models in Appendix C, where peer behavior affects achievement directly, leading to an additional indirect teacher effect, as captured in Duflo et al. (2011) and Jackson (2016).

Students choose behavior to maximize their expected utility from achievement net of the costs of behavior. To introduce a role for teaching practice in affecting behavior, we also permit that the marginal utility/cost of behavior varies with the practice, i.e.,

$$U_{it} = \gamma_y Y_{it} - \frac{\gamma_b}{2} b_{it}^2 + \gamma_{bp} P_j' b_{it}.$$

Student utility-maximizing behavior b_{it}^* is then

$$b_{it}^* = \frac{\gamma_y}{\gamma_b}(\beta_b + \beta_{by} Y_{it-1} + \beta_{b\bar{y}} m(Y_{-ic_t t-1})) + \frac{\gamma_{bp}}{\gamma_b} P_j'.$$

Behavior is increasing in initial achievement, peer initial achievement and, importantly, teaching practice. Classroom management practices may affect behavior directly through minimizing opportunities for disruptive behavior, whereas challenge/student-centered practices might do so by better engaging students in learning.

We cannot estimate (1) directly because we do not observe behavior. Instead, we assume that the achievement we observe in the data is coming through student optimizing behavior. To obtain an achievement production function we can take to the data, we plug in for utility-maximizing behavior

$$\begin{aligned}
Y_{it}^* = \tilde{\beta}_0 &+ (\beta_b \frac{\gamma_{bp}}{\gamma_b} + \beta_p) P_j' + (\beta_{b\bar{y}} \frac{\gamma_{bp}}{\gamma_b} + \beta_{p\bar{y}}) P_j' m(Y_{-ic_t t-1}) + \beta_{b\bar{y}}^2 \frac{\gamma_y}{\gamma_b} m(Y_{-ic_t t-1})^2 \\
&+ (2\beta_{b\bar{y}}\beta_b \frac{\gamma_y}{\gamma_b} + \beta_{\bar{y}}) m(Y_{-ic_t t-1}) + (\beta_y + 2\beta_{by}\beta_b \frac{\gamma_y}{\gamma_b}) Y_{it-1} + \beta_{by}^2 \frac{\gamma_y}{\gamma_b} Y_{it-1}^2 \\
&+ (\beta_{by} \frac{\gamma_{bp}}{\gamma_b} + \beta_{py}) P_j' Y_{it-1} + 2\beta_{by}\beta_{b\bar{y}} \frac{\gamma_y}{\gamma_b} Y_{it-1} m(Y_{-ic_t t-1}) + \epsilon_{it}, \\
=&\alpha_0 + \alpha_p P_j' + \alpha_{p\bar{y}} P_j' m(Y_{-ic_t t-1}) + \alpha_{\bar{y}} m(Y_{-ic_t t-1}) + \alpha_{\bar{y}2} m(Y_{-ic_t t-1})^2 + \\
&\alpha_y Y_{it-1} + \alpha_{y2} Y_{it-1}^2 + \alpha_{py} P_j' Y_{it-1} + \alpha_{y\bar{y}} Y_{it-1} m(Y_{-ic_t t-1}) + \epsilon_{it}.
\end{aligned} \tag{2}$$

Note that even if $\beta_p = \beta_{py} = \beta_{p\bar{y}} = 0$, so that teaching practice does not affect achievement directly and, more importantly, does not have direct complementarities with peer achievement, complementarities could still arise indirectly through the behavioral channel. This relies on two intuitive conditions. First, student behavior is affected by practice ($\beta_{bp} \neq 0$). Second, the benefits of peers varies with the student's behavior ($\beta_{b\bar{y}} \neq 0$). In Appendix C, we discuss some alternative forms of the behavioral model which could also underlie these complementarities, including popular conformity-style models (Brock and Durlauf, 2001; Epple and Romano, 2010) or the classic treatment of the classroom environment as a congestible public good (Lazear, 2001).

4 Estimation

Our empirical strategy focuses on estimation of the reduced form model described in equation (2). This relates most closely to models estimated in the literature. Implicitly, this assumes that observed achievement is a result of utlility-maximizing behavior on the part of students. We take as a starting point that $m(Y_{-ic_tt-1}) = \bar{Y}_{-ic_tt-1}$ and expand to include the IQR of the peer initial achievement distribution in the application, i.e.,

$$Y_{it} = \alpha_0 + \alpha_p P_{j\prime} + \alpha_{p\bar{y}} P_{j\prime} \bar{Y}_{-it-1} + \alpha_{\bar{y}} \bar{Y}_{-ic_tt-1} + \alpha_{\bar{y}2} \bar{Y}^2_{-ic_tt-1} + \alpha_y Y_{it-1} + \alpha_{y2} Y^2_{it-1} + \alpha_{py} P_{j\prime} Y_{it-1} + \alpha_{y\bar{y}} Y_{it-1} \bar{Y}_{-ic_tt-1} + \epsilon_{it}. \quad (3)$$

Our main parameter of interest is $\alpha_{p\bar{y}}$, which captures how the marginal benefits of teaching practices vary with the classroom composition. To simplify exposition, we ignore the role of other student and teacher observable characteristics though we include these additional controls in the analysis. Without loss of generality, we set $E(P_r) = E(\bar{Y}_{-ic_tt-1}) = E(Y_{it-1}) = 0$. Demeaning these variables aids in interpretation of the level terms, α_p, $\alpha_{\bar{y}}$ and α_y, in equation (4) by making them invariant to adding the interaction terms to the equation. The coefficients on the interactions are the same whether we normalize the variables or not.

As discussed above, a unique aspect of these data is that teachers are randomly assigned to classrooms within randomization blocks. However, even with random assignment of teachers to classrooms, several important endogeneity concerns remain. First, there is considerable non-compliance to the random assignment in the data. Largely, this was because assignments were made from preliminary rosters before school administrators had a good sense of who would be attending their school. Second, classroom composition may be endogenous as principals were not required to randomly assign students to classrooms. Third, teaching practice may still be endogenous even with random assignment because of measurement error. We discuss each of these issues in turn.

4.1 Non-compliance

Because the data include an indicator of the teacher that was randomly assigned to the student, we can use standard approaches for dealing with non-compliance, relying on the variation from the randomly-assigned teacher. These "intent-to-treat" estimates replace the observed teaching practice with the randomly-assigned teaching practice. Let P_r denote the teaching practice of the randomly-assigned teacher, indexed $r = r(i,t)$, then

$$Y_{it} = \alpha_0 + \alpha_p P_r\prime + \alpha_{p\bar{y}} P_r\prime \bar{Y}_{-ic_tt-1} + \alpha_{\bar{y}} \bar{Y}_{-ic_tt-1} + \alpha_{\bar{y}2} \bar{Y}^2_{-ic_tt-1} + \alpha_y Y_{it-1} + \alpha_{y2} Y^2_{it-1} + \alpha_{py} P_r\prime Y_{it-1} + \alpha_{y\bar{y}} Y_{it-1} \bar{Y}_{-it-1} + \alpha_b + \tilde{\epsilon}_{it}. \quad (4)$$

Because teachers are randomly assigned at the randomization-block level, we include randomization-block fixed effects, α_b, where $b = b(i,t)$ indexes randomization blocks. We show that our results are very similar when we instrument the observed with the randomly-assigned teacher's teaching practice, and so focus primarily on the more conservative intent-to-treat estimates for simplicity.

4.2 Endogeneity of classroom composition

Classroom composition could be endogenous for two reasons. First, the principals were not required to assign classroom composition randomly, though there was incentive to create comparable classrooms within randomization blocks to make the random assignment of teachers to either classroom palatable. Second, non-compliance by students could lead the classroom composition to be endogenous even after addressing non-compliance at the teacher-level.

The question is then whether we can identify $\alpha_{p\bar{y}}$ even though $\bar{Y}_{-ic_tt-1}$ is potentially endogenous. Nizalova and Murtazashvili (2014) show that interactions of the random treatment with endogenous characteristics are indeed exogenous. Bun and Harrison (2018) expand this and provide weaker conditions for identification.[22] The key assumption is that $(\bar{Y}_{-ic_tt-1}, \epsilon_{it})$ are jointly

[22] See also Annan and Schlenker (2015) and Di Falco et al. (2018) for other examples of

independent of P_r, conditional on other controls. This means that matching of students to peers based on unobservables does not vary with teaching practice. Thus, the main concern is about potential re-sorting of students after teachers are randomly assigned. We do not believe this poses a threat to identification for several reasons, which we discuss in Section 4.4. Furthermore, we can test the potential magnitude of this concern directly using the initially randomly-assigned peers, as we discuss in Section 5.1.

4.3 Endogeneity of teaching practice

Recall that we have multiple observations of teaching practice taken from video observations from multiple raters of the teacher both in the initial observational year and in the random assignment year to help deal with potential measurement error in teaching practice. As in Araujo et al. (2016), our preferred approach is to use $t-1$ measures to capture the teaching practice. This addresses two related concerns. First, video raters may have difficulty separating the teacher's practice from the students they are teaching. Second, if teachers change their practice in response to classroom composition, then teaching practice would no longer be exogenous, violating our key identifying assumptions.

Our main strategy relies on the most straightforward approach to measurement by taking simple averages of the measures of practice (P_{rt-1}). To clarify the potential effects of measurement error on our estimates, let the subscript k capture different observations of the teaching practice, i.e.,

$$P_{rkt-1} = P_r + u_{rkt-1}. \tag{5}$$

Substituting in the average measured practice for the true measures, we have

$$\begin{aligned} Y_{it} = \alpha_0 + \alpha_p P_{rt-1} + \alpha_{p\bar{y}} P_{rt-1} \bar{Y}_{-ic_t t-1} + \alpha_{\bar{y}} \bar{Y}_{-ic_t t-1} + \alpha_{\bar{y}2} \bar{Y}^2_{-ic_t t-1} + \alpha_y Y_{it-1} + \alpha_{y2} Y^2_{it-1} \\ + \alpha_{py} P_{rt-1} Y_{it-1} + \alpha_{y\bar{y}} Y_{it-1} \bar{Y}_{-ic_t t-1} + \alpha_b + \nu_{it}, \end{aligned}$$

applications of this argument.

where $\nu_{it} = \tilde{\epsilon}_{it} - \alpha_p \bar{u}_{rt-1} - \alpha_{p\bar{y}} \bar{u}_{rt-1} \bar{Y}_{-ic_t t-1} - \alpha_{py} \bar{u}_{rt-1} Y_{it-1}$. Note that as the number of observations of practice increases, $\bar{u}_{rt-1}$ goes toward 0, if u_{rkt} is mean independent of $u_{rk't}$ for $k \neq k'$. This is reasonable in our setting given the use of multiple trained raters to rate the same teacher, leading to arguably independent random draws of rater-related measurement error.[23]

We show results are robust to using principal component analysis to construct our measures (the primary approach we have seen applied in this literature) or factor models to extract the underlying teaching practice from multiple measures as in equation (5). We are also aware of the concern that simply including extracted factors in nonlinear models does not completely deal with measurement error. To deal with nonlinear measurement error, we adapt the method developed in Hausman et al. (1991) to deal with nonlinear errors in variables models to our setting where the nonlinearity takes the form of interactions. We describe this approach in detail in Appendix D.2. If anything these results imply that our estimates of the interactions are biased toward 0, which is typical of these types of models in the literature (Jaccard and Wan, 1995; Busemeyer and Jones, 1983).

To the extent that practice is time-varying, the focus on $t-1$ measures may understate the total effect of teaching practice. For time-varying practice, we can extract instead the common component from the correlation between time $t-1$ and t practices, which captures a persistent aspect of teaching practice. We discuss in Section 5.1 the findings when we instrument contemporaneous teaching practices with $t-1$ practices. These results show that if anything our estimation strategy provides conservative estimates of the interaction of practice with classroom composition.

4.4 Testing identifying assumption

We perform a number of tests to ensure that our key identifying assumption holds. First, we test random assignment of teachers based on observable characteristics of peers by separately regressing randomly-assigned teaching

[23]In earlier versions, we also tried controlling for rater fixed effects in measures of practice to account for any systematic rater differences and again results were very similar.

practice (based on $t-1$ averages) on different variables of classroom composition after controlling for randomization block fixed effects.[24] Appendix Table A.5 presents these balancing tests which show that teaching practice is not correlated with either of our measures of classroom composition, whether we use observed peers or initially-assigned peer. This also alleviates concerns about attrition from the sample based on the assigned teacher. Second, regressions of the randomly-assigned teaching practice on student-level covariates (Appendix table A.6) also suggest that random assignment of teachers held.

Third, though random assignment to classrooms is not needed for identification, Appendix Table A.6 also presents balancing tests which regress student characteristics on peer characteristics to see if there is evidence of matching in the data. Again, the balancing test generally supports that there is no matching of students (either using the observed or initially-assigned peers), suggesting that classroom composition does not appear to be endogenous, at least in terms of observables.[25] Fourth, that results of our balancing tests on classroom composition are similar in the initially-assigned and observed classroom compositions helps to alleviate concerns about students re-sorting after teachers were randomly assigned and/or differential attrition.

Again, while we do not need random assignment to classrooms to identify the interactions, the particular threat is if students re-sort to classrooms based on the randomly-assigned teacher. The above tests, while supportive of this in observable characteristics, cannot completely rule out re-sorting based on unobservable characteristics of the teacher or students. That said, we test the implications of this directly by replacing the observed peer characteristics with the initially-assigned peer characteristics in our regressions. We show that results are robust to this setting in Section 5.1.

[24] A similar approach is implemented in Kane et al. (2013).

[25] We find 3 out of 22 coefficients to be statistically significantly different from 0 at the 0.1 level, which is less than expected by chance.

5 Results

We begin by testing whether either of our measures of classroom composition moderate either of our measures of teaching practice. Treating the teaching practices separately offers the best chance for detecting effects, but ultimately we are also interested in checking the robustness to including them in the same regression in Section 6. Results for classroom management and challenge/student-centered practices are presented in Panels A and B respectively of Table 2. Odd columns only include average peer prior achievement interacted with teaching practice (in addition to student prior achievement), while even columns additionally control for classroom interquartile range and its interaction with teaching practice.[26] Columns (1) and (2) report ITT results (i.e. P_{jt-1} is replaced with P_{rt-1} as per equation (4)). Columns (3) and (4) report TT estimates where P_{jt-1} is instrumented with P_{rt-1}. All models control for the student gender, race/ethnicity and exceptionalities listed in Appendix Table A.3, CKT as a measure of teacher aptitude, prior achievement and peer characteristics (and their squared-terms) and all pairwise interactions of peer variables and prior achievement. All standard errors are clustered at the randomization block level, which is the level of randomization and the level of the fixed effects.[27]

Panel A shows that classrooms benefit more from higher average peer initial achievement when the teacher uses good classroom management practices, which is consistent with the mechanisms discussed in our model. For example, ITT and TT results show that a one standard deviation increase in classroom management increase test scores around 7.4% to 8.9% of a standard deviation when peer average prior year performance is one standard deviation above the mean. In contrast, the even columns show that the effectiveness of classroom management practices does not vary significantly with the IQR in classroom prior achievement. On the one hand, these results have the intuitive

[26]See footnote 20 for an explanation of why we include IQR in our specifications rather than standard deviation.

[27]Specifications that cluster at the class level have standard errors that are marginally smaller.

Table 2: Teaching Practice and Classroom Composition

	Random Teacher		IV Actual with Rand. Teacher		Random Teacher and Class	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Classroom Management	0.005	0.008	0.002	0.006	0.003	0.005
	(0.018)	(0.019)	(0.021)	(0.021)	(0.019)	(0.020)
C.M. × Math$_{t-1}$	0.011	0.011	0.011	0.011	0.015	0.014
	(0.013)	(0.012)	(0.013)	(0.012)	(0.013)	(0.012)
C.M. × Avg. Peer Math$_{t-1}$	0.079***	0.074***	0.089***	0.084***	0.056***	0.051**
	(0.021)	(0.025)	(0.023)	(0.030)	(0.020)	(0.022)
C.M. × IQR Peer Math$_{t-1}$		-0.017		-0.014		-0.017
		(0.019)		(0.023)		(0.018)
P-value (joint signif. of teaching practice)	0.001	0.000	0.000	0.000	0.018	0.007
First Stage F-Stat.†			84.4	42.9		
Panel B						
Challenge/Student-Centered	0.018	0.018	0.017	0.015	0.017	0.014
	(0.022)	(0.020)	(0.026)	(0.023)	(0.023)	(0.022)
C.S.C × Math$_{t-1}$	0.016	0.012	0.017	0.013	0.022*	0.020
	(0.012)	(0.012)	(0.013)	(0.013)	(0.013)	(0.012)
C.S.C × Avg Peer Math$_{t-1}$	0.044***	0.031**	0.050***	0.037**	0.035**	0.039**
	(0.016)	(0.014)	(0.019)	(0.017)	(0.016)	(0.015)
C.S.C. × IQR Peer Math$_{t-1}$		−0.053***		−0.058***		−0.037***
		(0.014)		(0.014)		(0.013)
P-value (joint signif. of teaching practice)	0.004	0.000	0.002	0.000	0.005	0.000
First Stage F-Statistic†			67.1	53.4		

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Sample size is 2632. Lagged teaching practices are used throughout; columns (5) and (6) control for characteristics of initially randomly assigned peers. Panel A and B correspond to different regressions with math as the dependent variable. These regressions include randomization block fixed effects and controls for the level and a squared term of prior math achievement and average peer prior achievement, as well as CKT and student characteristics listed in Table A.3. Even columns also include the IQR in peer prior achievement. Whenever peer variables are included we also include their square, and all pairwise interactions of peer variables and prior achievement. †Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test.

interpretation that a student cannot benefit from higher-achieving peers if the teacher does not have good classroom management practices, which would foster positive classroom behaviors. On the other hand, it could be expected that classroom management practices are more effective among low-achieving students. Our finding is consistent with the understanding that classroom management is also an important challenge in higher-achieving classrooms, though the sources of disengagement may be different from in lower-achieving classrooms (Shernoff et al., 2003).

Overall, we find that classroom management is a jointly significant predictor of math achievement at the 99% level in most specifications. This is driven by the interactions with the peers. The level effects of classroom management practices are not statistically significantly different from 0 and point estimates are small. Moreover, the interactions between classroom management and individual prior achievement are not statistically significantly different from 0.

Panel B shows results for challenge/student-centered practices. Generally, we find that classes with higher average initial achievement also benefit more for challenge/student-centered practices. However, the benefits of challenge/student-centered practices are smaller in classrooms with higher IQR in initial achievement. A standard deviation increase in this practice leads to a 5 to 6% reduction in achievement for classrooms that are a standard deviation above average IQR. Like in the case of classroom management, the level effects of challenge/student-centered practices are not statistically significantly different from 0 and neither are the interactions with initial achievement. Furthermore, joint tests also confirm that challenge/student-centered practices are jointly statistically significant predictors of achievement at the 99% confidence level, which appears to be primarily driven by heterogeneity in effects by classroom composition.

In summary, the findings in Table 2 indicate that teaching practices show significant complementarities with classroom characteristics, ranging in magnitude from 3% to 8.9% of a standard deviation increase in math, for a standard deviation increase in teaching practice in a class that is one standard deviation above the mean in prior performance. We view these estimates as sizable given that some of the larger estimates of a standard deviation increase in teacher value-added on math scores range from 0.11 to 0.16 (Chetty et al., 2014). The magnitudes also compare favorably to estimated effects of teacher experience. A standard finding in the literature is that the first two years of teacher experience, where experience effects are largest, increase student performance by only 0.06 of a standard deviation (Ladd and Thompson, 2008). The interactions with classroom composition appear to be important for the joint significance of the practices, which we test further below. Furthermore, that challenge/student-centered practices interact negatively with IQR whereas classroom management practices do not, supports the need to consider different dimensions of teaching practice, a point we also consider further below.

5.1 Robustness

Endogeneity of classroom composition. Given that teachers are randomly assigned to classrooms and that we focus on $t-1$ practices, a primary remaining endogeneity concern, as discussed in Section 4, is potential resorting of students to classrooms based on the teacher who is randomly assigned. Balancing tests reported in Section 4.4 already suggest that this is not the case, in that observable student and peer characteristics are not correlated with the randomly-assigned teacher's practice. However, given that we observe the students who were initially randomly assigned to the teacher, we can also test whether estimates of the interaction are systematically different if we replace actual peers with randomly-assigned peers, thus also addressing potential selection on unobservables. These estimates are reported in columns (5) and (6) of Table 2. Interactions between classroom composition and teaching practice are not statistically significantly different from their comparable estimates in columns (1) and (2), though smaller in magnitude. This is consistent with a slight downward bias in columns (5) and (6) generated from random measurement error in peers.

Streamlined specification. We also consider the extent to which controls are driving our results by contrasting estimates in columns (1) and (2) of Table 2 to estimates where we use minimal controls, dropping student and teacher quality controls and the quadratics in prior achievement and classroom composition along with interactions between prior achievement and classroom composition. In an experimental setting, you would expect results to be robust to dropping controls and indeed we find this to be the case. Results (available upon request) are very similar. In particular, coefficients on classroom composition interactions with teaching practice are within 0.005 of each other.

Main effects of practice and individual heterogeneity. Appendix Table A.7 tests whether classroom management and challenge/student-centered practices appear to matter for math achievement when we do not include interactions with classroom composition. Given the breadth of the measures, it is perhaps surprising that none of the specifications (in both panels) show that teaching practices play a statistically significant role in math performance. This is consistent with the findings in Garrett and Steinberg (2015), where the average of all FFT measures do not seem to have a direct impact on students' performance in their ITT and IV specifications.[28]

Interestingly, these specifications also indicate that interactions of student prior achievement with classroom management or challenge/student-centered practices are statistically significantly different from 0 at the 90% confidence level in ITT and IV specifications. This individual heterogeneity goes away when we control for interactions with average peer math, as indicated in Table 2. This suggests that failure to account for heterogeneity by classroom composition may lead us to overstate individual student heterogeneity.

F-tests in this streamlined specification show that the coefficients associated with these practices are in most specifications not jointly significant. Thus, allowing for heterogeneity by classroom composition is central to de-

[28]We also performed ITT specifications where average of all FFT measures are included as a regressor (instead of the subdomains). Our results also show that there is no direct impact of average FFT on student achievement. The negligible level effects of teaching practices are also not explained by measurement error.

tecting effects of these teaching practices on math.

Measurement error in teaching practice. An additional concern with our findings is to what extent measurement error in our key teaching practice variables affects our results (e.g. lack of significance in the level of the teaching practice measures). In order to address this point, we implement a measurement error correction strategy that follows Hausman et al. (1991). This approach is more convenient than the usual IV strategy that accounts for error in variables, because the variables of interest enter non-linearly into our model and we are over-identified by having more than 2 measures of each practice. In appendix D.2, we provide a description of how we adapt the Hausman et al. (1991) method to our context, and describe results obtained after implementing it. For completeness, we also report results when performing IV corrections (i.e. instrumenting one of the measures that corresponds to a given teaching practice with the remaining measures of that teaching practice). Overall, the findings indicate that our current strategy of taking averages of the teaching practice variables provides similar results to strategies that correct for measurement error following these alternative approaches. The level effects and interactions with initial achievement remain close to 0, but the interactions with classroom composition increase slightly after correcting for measurement error.

Sample selection. We also performed the same regressions as in Table 2 on a broader sample to show that our results are not driven by sample selection (see Appendix D.3).

Contemporaneous teaching practice. One implication of focusing on lagged measures of teaching practice is that our estimates of the interactions between classroom composition and teaching practice may understate the true effects. While we prefer focusing on these conservative estimates because of concerns about the endogeneity of contemporaneous teaching practice, we also explore how the interactions of teaching practice with classroom composition change when we instrument for contemporaneous teaching practices with lagged teaching practices. These results are presented in Appendix table A.9 and discussed in detail in Appendix D.4. We show that interactions between teaching practice and classroom composition remain robust, but (as expected) are significantly larger in magnitude.

Subdomains. We also consider whether our aggregation of the 8 subdomains into 2 separable components masks important heterogeneity. Appendix Table A.10 shows that similar findings hold when we consider the 8 different subdomains. The estimates, while still significant, are generally smaller in magnitude as would be expected due to increased measurement error.[29]

6 Mechanisms

6.1 Teaching Practice vs. Teacher "Quality"

While the previous section provides compelling evidence that teacher effectiveness varies by classroom composition, we now explore the extent to which classroom-management and challenge/student-centered practices may proxy for similar aspects of teacher effectiveness and/or whether more standard, unidimensional measures of teacher quality are the primary channel through which our teacher effectiveness measures operate. For instance, teachers who have better classroom management practices may also engage in more challenge/ student-centered practices; therefore not including both domains in the same specification may bias our estimates. This exploration raises a number of interesting questions. To be clear, there is no consensus on how teaching quality should be measured, and FFT was designed to capture different aspects of effective teaching. This means that in some ways classroom management and challenge/student-centered practices are in fact measures of quality. Furthermore, the fact that classroom-management and challenge/student-centered practices interact differently with classroom composition already suggests that

[29] See Appendix section D.5 for related discussion.

a single unidimensional quality may not be correct. Yet, we have other relevant unidimensional scales of quality, such as the Content Knowledge for Teaching assessment, as well as principal and student surveys, which we consider here.

In order to address these points, Table 3, Columns (1) and (2) present ITT (i.e. P_{jt-1} is replaced with P_{rt-1}) and IV (i.e. P_{jt-1} is instrumented with P_{rt-1}) results from a model that simultaneously controls for classroom management and challenge/student-centered practices and their interactions with peer composition. These results show that interactions of classroom management with the average peer initial achievement are robust, but seem to explain the interaction of challenge/student-centered practices with the average peer initial achievement in the previous tables because of strong correlations between these two practices. In contrast, interactions of challenge/student-centered practices with the IQR in peer initial achievement remain robust.[30]

Columns (3) to (5) of Table 3 report results from ITT specifications similar to column (1) where we additionally include different proxies for overall teacher "quality" and their interactions with classroom characteristics.[31] First, we include teacher performance in the Content Knowledge for Teaching (CKT) assessment interacted with classroom characteristics. Second, we include the teacher's lagged average score on student assessments from the TRIPOD survey. TRIPOD assesses the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth.[32] Finally, we include school principal evaluations of teachers performance which are reported in the MET database.[33] These results show that across all specifications our key interactions between teaching practices and classroom composition remain significant, and the size of these coefficients is very similar to our previous specifications. None of these alternative mea-

[30] Appendix tables A.11 report all the parameters of these specifications.

[31] Notice that in all previous specifications, we were controlling for a measure of teacher aptitude (i.e. CKT), but it was not interacted with classroom characteristics.

[32] Tripod is a protocol that measures teacher effectiveness based on student surveys. See Kane and Staiger (2012) for a description of this tool and the importance for predicting teacher value-added.

[33] The fact that our specifications include randomization blocks (which in this case are school-grade fixed effects) should account for systematic difference in principals' reporting.

Table 3: Teaching Practices and Alternative Teacher "Quality" Controls

	Random Teacher	IV Actual with Rand. Teacher	Random Teacher Alt. Teacher Control: CKT	7C	PSVY
	(1)	(2)	(3)	(4)	(5)
Classroom Management	−0.012	−0.016	−0.014	−0.016	−0.015
	(0.020)	(0.022)	(0.020)	(0.020)	(0.019)
C.M. × Math$_{t-1}$	0.004	0.004	0.011	0.004	0.003
	(0.020)	(0.021)	(0.019)	(0.019)	(0.019)
C.M. × Avg Peer Math$_{t-1}$	0.076**	0.087**	0.077**	0.076**	0.076***
	(0.029)	(0.036)	(0.030)	(0.029)	(0.027)
C.M.× IQR Peer Math$_{t-1}$	0.026	0.035	0.026	0.026	0.026
	(0.022)	(0.026)	(0.022)	(0.023)	(0.021)
Challenge/Student-Centered	0.026	0.025	0.026	0.026	0.011
	(0.023)	(0.025)	(0.022)	(0.022)	(0.024)
C.S.C. × Math$_{t-1}$	0.010	0.011	0.002	0.016	0.005
	(0.020)	(0.021)	(0.020)	(0.019)	(0.019)
C.S.C. × Avg Peer Math$_{t-1}$	−0.010	−0.009	−0.010	−0.010	−0.005
	(0.019)	(0.022)	(0.019)	(0.019)	(0.019)
C.S.C. × IQR Peer Math$_{t-1}$	−0.062***	−0.071***	−0.063***	−0.057**	−0.054**
	(0.017)	(0.019)	(0.017)	(0.021)	(0.021)
Alt. Teacher Control			−0.008	−0.006	0.055***
			(0.016)	(0.019)	(0.017)
T.C. × Math$_{t-1}$			0.044***	−0.029**	0.032**
			(0.014)	(0.013)	(0.013)
T.C. × Avg Peer Math$_{t-1}$			−0.019	−0.007	−0.016
			(0.018)	(0.020)	(0.016)
T.C. × IQR Peer Math$_{t-1}$			−0.012	−0.017	−0.003
			(0.021)	(0.021)	(0.016)
First Stage F-Statistic†		27.7			

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Sample size is 2632. Dependent variable is math and teaching practices are measured at $t-1$. Regressions include the same set of controls as in Table 2. † Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test. "TC" denotes the alternate teaching control in the column header; *CKT*, *7C* (overall Tripod student survey teacher ratings), and *PSVY* (principal assessments of teacher quality). See Appendix Table A.11 for all parameters.

sures of "quality" interact with peer average initial achievement and IQR in the same way as our two practices. In contrast to our practice measures, these show statistically significant heterogeneity in effects by the student's initial achievement, suggesting that "quality" as measured through CKT and principal assessments matters more for higher-achieving students.

We consider additional specifications that include teacher experience (as an alternative teacher control) and its interactions with classroom and students characteristics. These produce almost identical results to those reported in columns (3)-(5) of Table 3.[34] We also included teacher value-added from the previous year as an alternative measure of teacher quality. Again, the estimates of the interactions of teaching practice with classroom composition are very similar, providing further support that these measures are picking up something different from a unidimensional measure of quality.[35]

6.2 Overall FFT

Often FFT is treated as a unidimensional measure of quality. We also consider results when we include the average FFT of the teacher in the regressions, rather than classroom management and challenge/student-centered practices separately. Results (available upon request) are similar to those in column (1) of Table 3. We find that like classroom management, overall FFT interacts with the average classroom composition and is statistically significantly different from 0. The coefficient is slightly smaller, 0.06. Like challenge/student-centered practices, FFT negatively interacts with IQR of classroom prior achievement and is statistically significantly different from 0. Again the coefficient is slightly smaller in magnitude, -0.05. Also similarly, there is no evidence of a level effect or interaction with prior achievement. All these results are in line with our previous findings. It makes sense that they are slightly smaller in magnitude, given that the aggregate mixes the 2 practices. We prefer our spec-

[34]Results available upon request. This is true with both a continuous measure of experience and an indicator of whether the teacher has three or less years of experience.

[35]These results are available upon request. We do not emphasize these results because teacher value-added (i.e. adjusted random effects) models inherently neglect the presence of classroom-teacher interactions.

ification that treats the practices separately because it illustrates how not all aspects of FFT interact positively with the average of classroom composition or negatively with the IQR of prior achievement.

6.3 Class size

Because IQR is correlated with class size, an interesting question is whether interactions of challenge/ student-centered practices are driven by larger class sizes. We test this by adding interactions of classroom management and challenge/student-centered practices with class size to column (1) of Table 3. We do not include these results as we find no evidence that either practices interacts with class size. Furthermore, positive interactions of classroom management and average peer prior achievement and negative interactions of challenge/student-centered practices with the IQR remain robust, and if anything increase in magnitude with the additional controls.

7 Evaluating Teachers

Teacher evaluation protocols have taken center stage in many education policy debates in recent decades. For example, policymakers in the US have widely implemented accountability programs that intended to reward or punish teachers based on students gains in achievement.[36] More recently, schools have also incorporated classroom observation protocols like FFT to further assess teachers. For example, in 2012, the New Jersey Legislature passed the TEACHNJ Act, which mandated implementation of a new teacher evaluation system starting in the 2013 - 2014 school year and links tenure decisions to evaluation ratings. In response to this mandate, New Jersey has developed the program *AchieveNJ* that relies on classroom observation protocols such as FFT to evaluate teachers.[37]

[36] For example, North Carolina implemented the ABCs (Accountability for Basic skills and for local Control) program in 1997, and the US federal government developed the NCLB (No Child Left Behind) program in 2002.

[37] Teacher practice accounts for 55% of the teacher evaluations. The following link provides the list of approved teacher practice evaluation instruments used in New Jersey (see below):

Due to the increasing availability of data to evaluate teachers (e.g. value-added measures), scholars have highlighted the benefits of replacing the least effective 5% to 7% of teachers with average teachers (Chetty et al., 2014; Hanushek, 2011). However, these exercises in evaluating teachers and determining rankings rely crucially on the assumption that teacher effectiveness can be isolated from classroom characteristics. Our findings of a statistically significant complementarity between teaching practices and classroom composition suggest that the estimated teacher contributions will depend on classroom composition. To further quantify the relevance of these complementarities within the context of our empirical strategy, we implement simulation exercises that aim to determine how rankings of teacher contributions to learning vary when teachers are re-allocated into different classrooms.

We focus on our estimates in column (5) of Table 3 to create a measure for teacher contribution to learning based on our measures of teaching practices. We choose these estimates because they include a role for overall teacher quality through principal surveys, which is shown to have a direct effect on math achievement. We define the teacher contribution as

$$\hat{TC}_j = \frac{1}{N_j}\sum_{i=1}^{N_j}(\hat{\alpha}_p P_j\prime + \hat{\alpha}_{p\bar{y}} P_j\prime Y_{it-1} + \hat{\alpha}_{py} P_j\prime \bar{Y}_{c_t t-1} + \hat{\alpha}_{pI} P_j\prime IQR_{c_t t-1} + R_{VAj}),$$

where $IQR_{c_t t-1}$ denotes the interquantile range of the classes prior achievement, N_j denotes the class size for teacher j, α_{pI} the related coefficient estimate, P_j is a vector of classroom management practices, challenge/student-centered practices (using the average prior measures as discussed above), the principal's evaluation of teacher quality, and R_{VA_j} corresponds to estimates of "residual teacher value-added" that are recovered from the regression residuals.[38] We explore two different thought experiments to understand the mag-

https://www.nj.gov/education/AchieveNJ/teacher/approvedlist.pdf

[38]In particular, we construct these measures of teacher value-added by applying the shrinkage technique that is commonly used in the value-added literature (i.e. shrink the mean residuals by teacher from the specification in column (5) of Table 3 We name it "residual teacher value-added" because it represents only a part of the total teacher contribution to student learning).

nitude of the effects.

First, consider a teacher whose classroom management, challenge/student-centered practices, principal evaluation are 1 standard deviation above average and who has the average residual-value-added. Holding all other teachers classroom compositions fixed, but giving this teacher a classroom that is one standard deviation higher (relative to the mean) average classroom peer prior achievement (holding IQR fixed) would increase her teaching contribution to student learning ($\hat{TC}$) by 0.07 of a standard deviation and increase her rank in the teacher contribution distribution by 0.176 percentiles (on a 0 to 1 rank scale).[39] In a similar vein, holding the average peer initial achievement fixed and increasing IQR by one standard deviation would decrease teacher contribution by 0.03 and decrease her rank by 0.076 percentiles.

Because the above simulation makes the unrealistic assumption that we can change one teacher's classroom composition holding all other classroom compositions fixed, we also consider a simulation that reshuffles teachers within the randomization block.[40] This is potentially more similar to the usual allocation problem that principals face every year for a given math course/level, which is how the randomization blocks were defined. In this case, $\hat{TC}$ changes in absolute value by 0.30 of a standard deviation and the teacher's rank changes on average by 0.12 percentiles. This also leads to a shift in teacher relative rank in 22% of the randomization blocks. These counterfactual changes in rank are not trivial from a policy perspective. Our simulation shows that around 35% of the teachers that were ranked in the bottom 5 to 10% of the teaching contribution would no longer belong to that group after this local re-allocation of classrooms.

In summary, simulations suggest that complementarities between teaching practices and classroom composition play a key role in determining teacher contributions to learning, with sizable effects on teacher rank and in terms of

[39]The standard deviation of $\hat{TC}_j$ is 0.202. Moreover, note that 1 standard deviation average peer achievement (which is in standard deviation units) also increases prior student achievement by 0.42.

[40]If a randomization block has more than two classrooms, then teachers are reshuffled at random into a different classroom.

standard deviations of learning. In this regard, our findings suggest caution in implementing policies that aim to replace the bottom 5% of teachers given that teachers relative ranks are likely to depend on the characteristics of the classroom they are facing.

8 Conclusion

In this paper, we illustrate that the effects of teaching practice vary significantly with classroom composition. Our preferred estimates indicate that classroom management practices increase math achievement by 0.09 of a standard deviation when average classroom initial peer math performance is 1 standard deviation above average. In contrast, challenge/student-centered practices decrease math performance by -0.07 of a standard deviation when the classroom IQR in initial achievement is 1 standard deviation above average. We view these estimates as sizable given that some of the larger estimates of a standard deviation increase in teacher value-added, which is based on unobservable teacher contributions to math, range from 0.11 to 0.16 (Chetty et al., 2014). Simulations also illustrate that reassigning classrooms to teachers within randomization blocks would change their teacher contribution (in absolute value) by 0.30 standard deviations and change their rank on average by 0.12 percentiles.

We make four key contributions to the literature on teacher effectiveness. First, we illustrate that failure to account for moderating effects of classroom composition may lead researchers to severely misstate the importance of a given measured teaching practice for achievement. This helps address the common mystery of why teacher effectiveness is so hard to measure and may even help reconcile mixed findings in different contexts. Second, within-school changes in classroom composition lead to significant changes in teacher rank and predictions of teacher effectiveness based on math value-added.

Third, failure to account for the moderating effects of classroom composition also leads us to overstate the importance of individual student-level heterogeneity in the effects of teaching practice. Indeed, in our context, it appears that all heterogeneity is driven by classroom composition. Fourth, the focus in the literature on a single, unidimensional measure of teacher effectiveness may be misguided. Our two measures of teacher effectiveness interact with different aspects of classroom composition. Furthermore, we show that our estimated interactions of teaching practice with classroom composition remain after controlling for additional standard measures of teacher "quality," such as Content Knowledge for Teaching Assessment, student evaluations and principal surveys. In contrast to our FFT-based teaching practice measures, these additional measures of teacher quality show some evidence of heterogeneity by student prior achievement, but do not interact with classroom composition.

Our results have important implications for policies related to (1) teacher evaluation and accountability and (2) teacher professional development and training. Classroom observations of teaching practice–scored using the FFT and other protocols–are now routinely used in annual teacher evaluation and accountability. Our findings suggest that, depending on teachers' classroom context, specific domains of instructional practice may be more relevant to teacher effectiveness than others. As such, specific domains of instruction (rather than an overall observational score) may be emphasized in accountability systems depending on teaching assignments and/or school context. Moreover, this presents a challenge for measuring teacher effectiveness in a way that does not unfairly penalize teachers who teach more disadvantaged students or more heterogeneous classrooms.

In terms of teacher professional development and training, our findings reinforce the importance of explicit attention to challenges stemming from classroom-achievement heterogeneity (Cohen and Lotan, 1997; Seaton et al., 2010). We further find that scores on protocol subdomains do not appear to be as orthogonal in practice as they are in principle, or are intended to be.[41] Fur-

[41] That is, the MET observational protocol seem to have been developed as *formative* measures of instruction, where ideally the protocol would be useful in assessing "weak points" to target for instructional improvement. This is our own interpretation of these protocol. The supporting documentation we examined for the FFT protocol for example, does not specifically address the extent to which it was designed to measure a formative construct (Danielson, 2011, 2012).

ther research could benefit from determining how to more fully differentiate different aspects of teaching practice to make more formative recommendations for teacher training and development. That said, our research provides compelling evidence that any such recommendations should be adapted to the challenges faced by different school and classroom contexts.

Finally, our findings also have important implications for the peer effects literature. Because the effects of peers vary significantly with teaching practice, this suggests that failure to account for these interactions may also severely understate the importance of peers in different contexts. Furthermore, it suggests the potential for a change in policy emphasis from reallocating students to classrooms to meet different achievement objectives (which can be costly and involve severe tradeoffs among different types of students) to determining teaching practices that best fit different classroom contexts or better-matching of teachers to classrooms.

References

AIR, "Center on Great Teachers and Leaders: Databases on state teacher and principal evaluation policies," 2013.

Annan, Francis and Wolfram Schlenker, "Federal Crop Insurance and the Disincentive to Adapt to Extreme Heat," *The American Economic Review*, 2015, *105* (5), 262–266.

Araujo, Maria Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady, "Teacher Quality and Learning Outcomes in Kindergarten," *Quarterly Journal of Economics*, 2016, *131* (3).

Bacher-Hicks, Andrew, Mark J Chin, Thomas J Kane, and Douglas O Staiger, "An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys," 2017, (No. NBER 23478).

Brock, William A. and Steven N. Durlauf, “Interactions-Based Models,” in James Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 5, Amsterdam: Elsevier, 2001, pp. 3297–3380.

Bun, Maurice J. G. and Teresa D. Harrison, “OLS and IV estimation of regression models including endogenous interaction terms,” *Econometric Reviews*, 2018, *Forthcoming*, 1–14.

Busemeyer, Jerome R. and Lawrence E. Jones, “Analysis of multiplicative combination rules when the causal variables are measured with error,” *Psychological Bulletin*, 1983, *93* (3), 549–562.

Cantrell, Steve and Thomas J Kane, “Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project’s three-year study,” *Policy and Practice Brief*, 2013.

Chetty, Raj, John N Friedman, and Jonah E Rockoff, “Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates,” *The American Economic Review*, 2014, *104* (9), 2593–2632.

Cohen, E. G. and R. A. Lotan, *Working for Equity in Heterogeneous Classrooms: Sociological Theory in Practice*, New York: Sociology of Education Series. Teachers College Press, 1234 Amsterdam Avenue, , NY 10027 (paperback: ISBN-0-8077-3643-0; clothbound: ISBN-0-8077-3644-9, 1997.

Connor, Carol McDonald, Frederick J Morrison, and Leslie E Katch, “Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading,” *Scientific studies of reading*, 2004, *8* (4), 305–336.

Danielson, Charlotte, “The framework for teaching evaluation instrument,” The Danielson Group Princeton, NJ 2011.

_ , “Teacher evaluation: What’s fair? What’s effective?,” *Educational Leadership*, 2012, *70* (3), 32–37.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 2011, *101* (5), 1739–1774.

Epple, Dennis and Richard Romano, "Peer Effects in Education: A Survey of the Theory and Evidence," in Jess Benhabib, Alberto Bisin, and Matthew O. Jackson, eds., *Handbook of Social Economics*, Vol. 1B, Amsterdam, The Netherlands: North-Holland, 2010, chapter 20, pp. 1053–1164.

Falco, Salvatore Di, Francesco Feri, Paolo Pin, and Xavier Vollenweider, "Ties that bind: Network redistributive pressure and economic decisions in village economies," *Journal of Development Economics*, 2018, *131*, 123–131.

Fox, Lindsay, "Playing to Teachers' Strengths: Using multiple measures of teacher effectiveness to improve teacher assignments," *Education Finance and Policy*, 2016.

Gamoran, Adam, Walter G. Secada, and Corab Marrett, "The organizational context of teaching and learning: Changing theoretical perspectives," in M. Hallinan, ed., *Handbook of the Sociology of Education*, New York: Kluwer Academic/Plenum, 2000.

Garrett, Rachel and Matthew P Steinberg, "Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students," *Educational Evaluation and Policy Analysis*, 2015, *37* (2), 224–242.

Guryan, Jonathan, Kory Kroft, and Matthew J. Notowidigdo, "Peer Effects in the Workplace: Evidence from Random Groupings in Professional Golf Tournaments," *American Economic Journal: Applied Economics*, October 2009, *1* (4), 34–68.

Hamilton, L. S., "Measuring teaching quality using student achievement tests: Lessons from educators' responses to No Child Left Behind," in Sean

Kelly, ed., *Assessing teacher quality: Understanding teacher effects on instruction and achievement*, New York: Teachers College Press, 2012.

Hanushek, Eric A, "How much is a good teacher worth," *Education next*, 2011, *11* (3), 40–45.

Hausman, Jerry A., Whitney K. Newey, Hidehiko Ichimura, and James L. Powell, "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 1991, *50* (3), 273 – 295.

Jaccard, James and Choi K Wan, "Measurement Error in the Analysis of Interaction Effects Between Continuous Predictors Using Multiple Regression: Multiple Indicator and Structural Equation Approaches," *Quantitative Methods in Psychology*, 1995, *117* (2), 348–357.

Jackson, C. Kirabo, "Match quality, worker productivity and worker mobility: direct evidence from teachers," *The review of economis and statistics*, October 2013, *95* (4), 1096–1116.

_ , "The Effect of Single-Sex Education on Test Scores, School Completion, Arrests, and Teen Motherhood: Evidence from School Transitions," 2016. NBER working paper 22222.

_ , Jonah E. Rockoff, and Douglas O. Staiger, "Teacher Effects and Teacher-Related Policies," *Annual Review of Economics*, 2014, *6* (1), 801–825.

Kane, Thomas J and Douglas O Staiger, "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project.," *Bill & Melinda Gates Foundation*, 2012.

_ , Daniel F McCaffrey, Trey Miller, and Douglas O Staiger, "Have we identified effective teachers? Validating measures of effective teaching using random assignment," in "Research Paper. MET Project. Bill & Melinda Gates Foundation" Citeseer 2013.

Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten, "Identifying Effective Classroom Practices Using Student Achievement Data," *Journal of Human Resources*, 2011, *46* (3), 587–613.

Kennedy, M. M., "Introduction: The Uncertain Relationship between Teacher Assessment and Teacher Quality," in "Teacher Assessment and the Quest for Teacher Quality," San francisco: Jossey Bass, 2010.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff, "Value-added modeling: A review," *Economics of Education Review*, 2015, *47*, 180 – 195.

Kolenikov, Stanislav, "Confirmatory Factor Analysis Using Confa," *The Stata Journal*, 2009, *9* (3), 329–373.

Konstantopoulos, Spyros, "Effects of Teachers on Minority and Disadvantaged Students' Achievement in the Early Grades," *The Elementary School Journal*, 2009, *110* (1), 92–113.

Ladd, Helen and Edgar Thompson, "Teacher effects: What do we know," 01 2008, *21.*

Lavy, Victor, "What Makes an Effective Teacher? Quasi-Experimental Evidence," *CESifo Economic Studies*, 2015.

_ , M. Daniele Paserman, and Analia Schlosser, "Inside the Black Box of Ability Peer Effects: Evidence from Variation in the Proportion of Low Achievers in the Classroom," *Economic Journal*, 03 2012, *122* (559), 208–237.

Lazear, Edward P., "Educational production," *Quarterly Journal of Economics*, 2001, *116* (3), 777–803.

Lee, Soohyung, Lesley J. Turner, Seokjin Woo, and Kyunghee Kim, "All or Nothing? The Impact of School and Classroom Gender Composition on Effort and Academic Achievement," 2014. NBER working paper 20722.

Manski, Charles, "Identification of endogenous social effects: The reflection problem," *The Review of Economic Studies*, 1993, *60* (3), 531–542.

MET project, "Danielson's framework for teaching for classroom observations," Bill and Melinda Gates foundation October 2010.

Mihaly, Kata, Daniel F. McCaffrey, Douglas Staiger, and J.R. Lockwood, "A composite estimator of effective teaching," *MET Project Research Paper, Bill & Melinda Gates Foundation*, 2013.

Nizalova, Olena Y. and Irina Murtazashvili, "Exogenous Treatment and Endogenous Factors: Vanishing of Omitted Variable Bias on the Interaction Term," *Journal of Econometric Methods*, 2014, *5* (1), 71–77.

Rivkin, Steven G., Eric A. Hanushek, and John F. Kain, "Teachers, Schools, and Academic Achievement," *Econometrica*, 03 2005, *73* (2), 417–458.

Rothstein, J., "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics*, February 2010, *125* (1), 175–214.

Sacerdote, Bruce, "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," in Erik Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of the Economics of Education*, Vol. 3, Elsevier, June 2011, chapter 4, pp. 249–277.

Seaton, Marjorie, Herbert W. Marsh, and Rhonda G. Craven, "Big-Fish-Little-Pond Effect: Generalizability and Moderation—Two Sides of the Same Coin," *American Educational Research Journal*, 2010, *47* (2), 390–433.

Shernoff, David J., Mihaly Csikszentmihalyi, Barbara Schneider, and Elisa Steele Shernoff, "Student Engagement in High School Classrooms from the Perspective of Flow Theory," *School Psychology Quarterly*, 2003, *18* (2), 158–176.

Stacy, Brian, Cassandra Guarino, Mark Reckase, and Jeffrey Wooldridge, "Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve?," Education Policy Center, Michigan State University, Working paper no. 35 2013.

Taylor, Eric S., "Skills, Job Tasks, and Productivity in Teaching: Evidence from a Randomized Trial of Instruction Practices," *Journal of Labor Economics*, 2018, *36* (3).

Zimmer, Ron, "A new twist in the education tracking debate," *Economics of Education Review*, 2003, *22* (3), 307–315.

For Online Publication

A Randomization

When schools joined the MET study in 2009-2010, principals were asked to identify groups of teachers that 1) were teaching the same subject to students in the same grade 2) were certified to teach common classes and 3) were expected to teach the same subject to students in the same grade the following year. These groups of teachers were called "exchange groups" The plan was for principals to create class rosters as similar as possible within an exchange group, and then send these rosters to MET to be randomly assigned to "exchangeable" teachers. One issue in practice was that when it came time to perform the randomization, not all teachers within an exchange group were able to teach during a common period. As a result, randomization was performed within subsets of exchange groups called "randomization blocks." In summary, MET requested scheduling information for 2,462 teachers from 865 exchange groups in 316 schools. From this, they created 668 randomization blocks from 619 exchange groups in 284 participating schools. The drop off in teachers can be attributed to either a school refusing to permit randomly swapping rosters, or all remaining MET project teachers leaving the school or the study prior to randomization. From these randomization blocks, 1,591 teachers were randomly assigned to class rosters. Teachers were lost either because they were not scheduled to teach the exchange group subject and grade level in 2010-2011 or they decided not to participate (Kane et al., 2013).[42]

Since assignments were made based on preliminary rosters at the end of the previous school year, before school administrators knew who would be attending their school, there was both attrition from the sample and additional students who moved into the school and needed to be incorporated in the sample. As a result, our analysis does not rely on the assumption that the

[42]The number of randomized teachers includes 386 high school teachers and 24 teachers from grades 4-8 for whom rosters were later found to be invalid by MET. We do not include these in our sample.

observed classroom composition is random, but rather exploits what we know to be random—the initial random assignment of teachers to classrooms. We discuss this further in Section 4. We cannot include students who were not in the randomization sample in our main analysis, which relies on the randomization, but we do include them as part of the calculation of classroom composition when prior test scores are available. For the average student in our final sample, 78% of classroom peers were included in randomization, and we observe prior test scores for 91% of classroom peers.

B Sample

Before restricting the sample, we calculate peer variables based on the largest available set of prior student test score that were assigned to a given student's classroom. Then, we restrict the sample in several ways necessary for the analysis, removing students who are missing either a contemporaneous or lagged test scores, or missing teaching practice or CKT (our measure of teacher quality) for the student's actual or randomly-assigned teacher. This brings the sample down from 5,730 to 4,201. We also drop the bottom percentile of class sizes (classrooms less than 7 students) out of concern for measurement error in classroom composition, leaving 4,121 students. Our identification strategy also requires at least 2 teachers per randomization block, the level of randomization. This restricts the sample further to 3,618 student. We also restrict to randomization blocks with at least a 50% compliance rate so that the randomly assigned teacher has some significant relationship with the assignment, restricting to 2,682 students. 50 additional observations are dropped because of duplication between classes and rechecking after these restrictions that the class size and teachers per randomization block criteria were met, leaving 2,632 students in the final sample.[43]

[43] We estimate our main results on the largest possible sample as well, after we remove the duplications, class size, teachers per randomization block restrictions and drop the compliance rate to 25 percent. We find results are not statistically significantly different, though point estimates are marginally smaller and standard errors are larger, as you would expect with lower compliance rates and noisier measures of classroom composition.

Table A.3 reports summary statistics for characteristics of the students in our final sample. This is a racially-diverse sample; 31% of students are black, 25% are white, 29% are Hispanic, and 11% are Asian, indicating that the school districts included in our data are not necessarily representative of the whole US population of students. Though the randomization sample was not selected to be representative, it is still useful to compare our estimation sample to the randomization sample (Appendix Table A.4). The samples are demographically very similar, and math scores of the selected sample are marginally lower, by about 0.1 of a standard deviation.

C Alternative Models

While the behavioral model in Section 3 posits some possible channels of complementarities, alternative plausible models of student behavior would produce similar complementarities. For instance, it is straightforward to add to the model that students conform to the average behavior of classmates, so that utility is

$$U_{it} = \gamma_y Y_{it} - \frac{\gamma_b}{2}(b_{it} - \gamma_{\bar{b}}\bar{b}_{-it})^2 + \gamma_{bp} P_j' b_{it}.$$

This captures the conformity-type peer effects that are the focus of the social interactions literature (Brock and Durlauf, 2001; Epple and Romano, 2010). In this case, optimal behavior would be a function of peer behavior and teaching practice and similar results would follow, except here the benefits of the teaching practice are amplified through the re-enforcing behavior of peers. For instance, a teacher's classroom management practice encourages a student and her peers to behave better, and the better behavior of peers further encourages the student's own better behavior and vice-versa. The interaction between teaching practice and peer initial achievement would follow again in this model because the marginal product of good behavior differs with peer initial achievement.

Furthermore, we could also motivate the interaction between teachers and peers as arising through a production function that has complementarities between average peer behavior and own behavior, i.e.,

$$\begin{aligned}Y_{it} &= \beta_0 + \beta_b b_{it} + \beta_{by} b_{it} Y_{it-1} + \beta_{b\bar{y}} b_{it} m(Y_{-ic_t t-1}) + \beta_{b\bar{b}} b_{it} \bar{b}_{-it} + \beta_{\bar{b}} \bar{b}_{-it} \\ &+ \beta_y Y_{it-1} + \beta_{\bar{y}} m(Y_{-ic_t t-1}) + \beta_p P'_j + \beta_{py} P'_j Y_{it-1} + \beta_{p\bar{y}} P'_j m(Y_{-ic_t t-1}) + \epsilon_{it},\end{aligned}$$

where there are direct spillovers from peer behavior and the achievement benefits of behavior are increasing in peer behavior. This channel connects well with Lazear (2001)'s classic treatment of the classroom learning environment as a public good that is disrupted by student behaviors. The reduced form in this setting would be similar in structure to the above, when $m(Y_{-ic_t t-1}) = \bar{Y}_{-ic_t t-1}$, with the addition of the P_j^2 term arising through the interaction of own and peer behavior, both of which are increasing in P_j.

D Robustness

D.1 Direct Effects of Teaching Practices

Panels A and B of Table A.7 display estimates of the effect of classroom management and challenge/ student-centered practices, respectively on math performance. Even columns allow the effect of teaching practice to vary by a student's initial achievement. Results in columns (1) and (2) are naive OLS specifications, where the lagged teaching practice of the current teacher (P_{jt-1}) is the variable of interest. Columns (3) and (4) report intent-to-treat (ITT) estimates, replacing P_{jt-1} with the teaching practice of the randomly-assigned teacher (P_{rt-1}). Columns (5) and (6) present treatment on the treated (TT) results where P_{jt-1} is instrumented with P_{rt-1}. Across the board the level effects of the teaching practices are not statistically significantly different from 0, whereas some of the interactions with initial achievement are.

D.2 Nonlinear Measurement Error

To show how Hausman et al. (1991) can be adapted to our setting to deal with measurement error in teaching practice, we consider a simplified version of our main estimating equation (4). Results are robust to using this streamlined specification. Let $\tilde{Y}$ denotes Y demeaned at the randomization block level and similarly for other variables, then

$$\tilde{Y}_{it} = \alpha_p \tilde{P}_r + \alpha_{p\bar{y}} \widetilde{P_r \bar{Y}}_{-ic_t t-1} + \alpha_{\bar{y}} \tilde{\bar{Y}}_{-ic_t t-1} + \alpha_y \tilde{Y}_{it-1} + \alpha_{py} \widetilde{P_r Y}_{it-1} + \tilde{\epsilon}_{it}. \quad (6)$$

Thus, all variables are in deviations from the randomization block mean to capture the variation at the level of random assignment. Recall that P_r is the true practice, but it is measured with error. We adapt Hausman et al. (1991) in two ways. First, we relax the assumptions on the measurement model because we have more than 2 measures for each practice. Second, we adapt their approach for dealing with nonlinearities arising from polynomials in the variables of interest (that are measured with error), to our setting where nonlinearities arise from interactions of the teaching practices with controls.

The parameters of equation (6) are identified from

$$\begin{aligned}
E(\tilde{Y}_{it}) =& \alpha_p E(\tilde{P}_r) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}) + \alpha_y E(\tilde{Y}_{it-1}) && (7)\\
&+ \alpha_{py} E(\widetilde{P_r Y}_{it-1}) && (8)\\
E(\tilde{Y}_{it}\tilde{P}_r) =& \alpha_p E(\tilde{P}_r\tilde{P}_r) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}\tilde{P}_r) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}\tilde{P}_r) && (9)\\
&+ \alpha_y E(\tilde{Y}_{it-1}\tilde{P}_r) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{P}_r)\\
E(\tilde{Y}_{it}\tilde{\bar{Y}}_{-ic_t t-1}) =& \alpha_p E(\tilde{P}_r\tilde{\bar{Y}}_{-ic_t t-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}\tilde{\bar{Y}}_{-ic_t t-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}\tilde{\bar{Y}}_{-ic_t t-1}) && (10)\\
&+ \alpha_y E(\tilde{Y}_{it-1}\tilde{\bar{Y}}_{-ic_t t-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{\bar{Y}}_{-ic_t t-1}) && (11)\\
E(\tilde{Y}_{it}\widetilde{P_r \bar{Y}}_{-ic_t t-1}) =& \alpha_p E(\tilde{P}_r\widetilde{P_r \bar{Y}}_{-ic_t t-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}\widetilde{P_r \bar{Y}}_{-ic_t t-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}\widetilde{P_r \bar{Y}}_{-ic_t t-1})\\
&+ \alpha_y E(\tilde{Y}_{it-1}\widetilde{P_r \bar{Y}}_{-ic_t t-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\widetilde{P_r \bar{Y}}_{-ic_t t-1})\\
E(\tilde{Y}_{it}\tilde{Y}_{it-1}) =& \alpha_p E(\tilde{P}_r\tilde{Y}_{it-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}\tilde{Y}_{it-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}\tilde{Y}_{it-1})\\
&+ \alpha_y E(\tilde{Y}_{it-1}\tilde{Y}_{it-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{Y}_{it-1})\\
E(\tilde{Y}_{it}\widetilde{P_r Y}_{it-1}) =& \alpha_p E(\tilde{P}_r\widetilde{P_r Y}_{it-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ic_t t-1}\widetilde{P_r Y}_{it-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ic_t t-1}\widetilde{P_r Y}_{it-1})\\
&+ \alpha_y E(\tilde{Y}_{it-1}\widetilde{P_r Y}_{it-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\widetilde{P_r Y}_{it-1})
\end{aligned}$$

We need to recover all of the moments containing P_r. The issue is that P_r is not observed, so next we discuss how to use our measures of practice to recover these moments.

We assume that we have at least 3 demeaned measures of practice following equation 5, such that

$$P_{jkt} = \delta_k P_j + u_{jkt},$$

where $k = \{1, ..., K\}$ and $K \geq 3$. We focus the measurement equation around the mean reports for each subdomain, calculated over multiple videos and video raters, though we could apply adjustments to the individual level observations as well. Then, applying a normalization, $\delta_1 = 1$, we have

$$\frac{Cov(P_{jnt}, P_{jmt})}{Cov(P_{jnt}, P_{j1t})} = \frac{\delta_n \delta_m V(P_j)}{\delta_n V(P_j)} = \delta_m,$$

for $n, m \neq 1$ and $n \neq m$, thus permitting us to recover the parameters $\delta_2, ..., \delta_k$. Notice further that

$$E(P_{j1t}P_{jnt}) = \delta_n E(P_j^2), \text{ for } n \neq 1$$

and $E(P_j^2)$ is thus identified and similarly,

$$E(\tilde{P}_{j1t}\tilde{P}_{jnt}) = \delta_n E(\tilde{P}_j^2), \text{ for } n \neq 1,$$

given that measurement error is also uncorrelated across measures after removing randomization block fixed effects. Note that $E(\tilde{P}_j) = 0$.

We can use our anchor measure then to recover

$$
\begin{aligned}
E(\widetilde{P_{r1t}\bar{Y}}_{-ic_tt-1}) &= E(\widetilde{P_r\bar{Y}}_{-ic_tt-1}) \\
E(\widetilde{P_{r1t}Y}_{it-1}) &= E(\widetilde{P_rY}_{it-1}) \\
E(\tilde{Y}_{it}\tilde{P}_{r1t}) &= E(\tilde{Y}_{it}\tilde{P}_r) \\
E(\tilde{Y}_{it}\widetilde{P_{r1t}Y}_{it-1}) &= E(\tilde{Y}_{it}\widetilde{P_rY}_{it-1}) \\
E(\tilde{Y}_{it}\widetilde{P_{r1t}\bar{Y}}_{-ic_tt-1}) &= E(\tilde{Y}_{it}\widetilde{P_r\bar{Y}}_{-ic_tt-1})
\end{aligned}
$$

But to recover terms which have higher order products of P_r such as $E(\widetilde{P_rY}_{it-1}\tilde{P}_r)$ we rely on the ratio of covariances to first recover δ_2. We can then use our anchor measure and measurement two to recover

$$
E\left(\frac{\widetilde{P_1Y}_{it-1}\tilde{P}_2}{\delta_2}\right) = E(\widetilde{P_rY}_{it-1}\tilde{P}_r)
$$

Specifically, in estimation we pick an anchor measurement, P_1, and use it to construct the terms in equation (6). To construct rows two, four and six in the system (7), we multiply equation (6) by $\frac{\tilde{P}_{r2t}}{\delta_2}$, $\frac{\widetilde{P_{r2t}\bar{Y}}_{-ic_tt-1}}{\delta_2}$ and $\frac{\widetilde{P_{r2t}Y}_{it-1}}{\delta_2}$ and then take expectations. Note that we use measurement two when multiplying through and then divide by the measurement parameter we have recovered.

Estimation of the parameters from these moments is then straightforward. We recover the relevant moments from the measurement model and then plug them into the system defined in 7 and solve this system for the structural parameters. We can bootstrap standard errors, clustering at the randomization block level. Note that because we are overidentified, we can also test robustness to using different measures as our anchor.

Appendix Table A.8 shows results when we correct for measurement error by following two strategies. First, we present findings when we implement the Hausman et al. (1991) method described above, but we also report (for completeness) specifications when we instrument a given measure of a teaching practice at $t-1$ (e.g. creating an environment of respect and rapport when considering the broad category classroom management) with the remaining teaching practices at $t-1$ (e.g. managing student behaviors and classroom procedures). Overall, results indicate that taking averages across measurements that correspond to a specific broad teaching practice (i.e. classroom management or challenge/student-centered) lead to similar results to when we correct for measurement error by following other methods.

D.3 Sample Selection

We also performed similar regressions to Table 2 on a broader sample to test to what extent our sample restrictions described in Appendix B could drive our results. If instead we restrict the data to: a) fourth and fifth grade students who were randomly assigned a teacher, b) both a student's actual and randomly assigned teacher have non-missing year 1 teaching practice measures, c) students have non-missing test scores in both years, and d) keep randomization blocks with at least a 25% compliance rate, leads to a sample of 4086 students. This larger sample provides very similar (though noisier as would be expected) results to the main results in Table 2.

D.4 Contemporaneous teaching practice

Table A.9 shows robustness checks on our ITT results when we use lagged teaching practice to instrument for contemporaneous teaching practice focusing on our key interactions–challenge/student-centered practices interacted with the IQR in initial peer achievement and classroom management interacted with average initial peer achievement. It is difficult to instrument for all the entries of teaching practice and its interactions without running into a weak instrument problem. As a result, we build the argument sequentially to show that weak instruments are not driving estimates of our key interactions. Panel A shows results for classroom management and panel B for challenge/student-centered practices. The first column shows results for the

ITT when P_{rt-1}, $P_{rt-1}Y_{it-1}$, $P_{rt-1}\bar{Y}_{-ic_tt-1}$, $P_{rt-1}IQR_{c_tt-1}$ are all included in the regression. Then, column (2) shows that estimates of key interactions are robust when all other teaching practice terms are dropped except our main interactions of interest, i.e., $P_{rt-1}IQR_{c_tt-1}$ for challenge/student-centered and $P_{rt-1}\bar{Y}_{-ic_tt-1}$ for classroom management. Column (3) then instruments for $P_{rt}\bar{Y}_{-ic_tt-1}$ with $P_{rt-1}\bar{Y}_{-ic_tt-1}$ for classroom management and $P_{rt}IQR_{c_tt-1}$ with $P_{rt-1}IQR_{c_tt-1}$ for challenge/student-centered. The F-statistics for weak instrument tests are in both cases are 28 and 27 respectively, indicating that there is not a weak instrument problem. And, in both cases the estimated interactions are significantly larger, increasing from 0.08 to 0.22 for the case of classroom management with the average and -0.06 to -0.18 for student-centered practices with IQR.

Column (4) shows another variation of this when we continue to control for P_{rt-1}, $P_{rt-1}Y_{it-1}$, but only drop from the regression the irrelevant peer interactions, i.e., the interactions with IQR for classroom management and average initial peer achievement for challenge/student-centered. Column (5) controls for contemporaneous teaching practice in levels and interacted with prior achievement (P_{rt}, $P_{rt}Y_{it-1}$) and only instruments for key interactions of contemporaneous teaching practice with peer variables ($P_{rt}\bar{Y}_{-ic_tt-1}$ with $P_{rt-1}\bar{Y}_{-ic_tt-1}$ for classroom management and $P_{rt}IQR_{c_tt-1}$ with $P_{rt-1}IQR_{c_tt-1}$ for challenge/student-centered). Again, F-statistics for the weak instrument test are in all cases above 20 and the key variables of interest remain very similar to estimates in column (3) that do not control for level effects or interactions with initial achievement. Finally, column (6) instruments for all entries of contemporaneous teaching practice (i.e., P_{rt}, $P_{rt}Y_{it-1}$ are also instrumented with P_{rt-1} and $P_{rt-1}Y_{it-1}$), along with the key classroom composition interactions as in column (5). In this case, F-statistics on tests for weak instruments drop below 10, but we see that the estimated interactions with classroom composition remain remarkably stable, suggesting that the sizable effects are not driven by weak instruments.

D.5 Choosing Practices that Matter

A tension in using our composite measures of teaching practice is that they do not provide as fine-grained prescriptive evidence as desirable on what practices matter most in different settings, which arguably is consistent with the formative underpinnings of the FFT with its eight separate subdomains. With this in mind, we present in Table A.10 results at the subdomain level in order to complement the evidence from the aggregated subdomains, particularly mirroring results in column (2) of Table 2 for each subdomain separately. We offer two notes of caution when interpreting these results. First, a higher degree of measurement error should bias interactions toward zero. Second, the subdomains are highly correlated as revealed by the exploratory factor model.

A main pattern we see in this table is that there is a positive interaction with average peer prior achievement with all the subdomains that aggregate to make up classroom management (the first 3 columns of Table 6, Panel A), i.e., creating an environment of respect and rapport (CERR), managing classroom procedure (MCP) and managing student behaviors (MSB).[44] Each of these subdomains shows a positive interaction with average peer achievement. Managing student behavior (MSB) is the largest, but not statistically significantly different from the other subdomains. Teachers with high levels of MSB are characterized by establishing clear expectations for student conduct and by implementing them efficiently. This suggests that peer effects are amplified by teachers that can preempt misbehavior in the classroom. The two other subdomains, MCP and CERR, are linked to teachers' skills in managing more general aspects of the classroom environment, including instructional groups, transitions and teacher/student interactions. The significant positive interactions with average prior achievement suggests that there are a number of interrelated practices beyond just limiting disruptive behaviors, which create an environment where students can benefit more from having higher-achieving peers.

Second, across the board the five subdomains which make up challenge/student-

[44]See Table A.1 in the appendix for the definition of each subdomain.

centered practice exhibit negative interactions with class IQR. These include establishing a culture of learning (ECL), engaging students in learning (ESL), using questioning and discussion techniques (USDT), using assessment in instruction (UAI) and communicating with students (CS). Among these, communicating with students has the largest negative coefficient but also the highest standard error. The definition of these rubrics are closely related to promoting student active participation in the class as a key element of the learning process. More detailed consideration of the rubrics also reveals significant emphasis on challenging students in the different subdomains. Our findings indicate that the benefit of these practices are largely dependent on the heterogeneity in classroom prior achievement. Basically, promoting discussion among students may not constitute a good learning tool when *all* students cannot share a somewhat similar level of understanding on key concepts. For example, large heterogeneity in classroom achievement is likely to require different levels of complexity in the discussion, making the learning process more complicated. Likewise, it may be difficult to challenge all students when there there is a great deal of heterogeneity in background.

E Appendix Tables

Table A.1: Description of Framework for Teaching (FFT)

Classroom Management Practices	
Managing student behaviors (MSB)	Monitoring of student behavior, response to student misbehavior, expectations
Managing classroom procedures (MCP)	Management of instructional groups, transitions, and materials and supplies
Creating an environment of respect and rapport (CERR)	Teacher interactions with students and student interactions with each other
Challenge/Student-Centered Practices	
Establishing a culture of learning (ECL)	Importance of content and expectations for learning and achievement
Communicating with students (CS)	Expectations for learning, directions and procedures, explanations of content, use of oral and written language
Engaging students in learning (ESL)	Activities and assignments, grouping of students, instructional materials and resources, structure and pacing
Using assessment in instruction (UAI)	Assessment criteria, monitoring of student learning, feedback to students, student self-assessment and monitoring of progress
Using questioning and discussion techniques (USDT)	Quality of questions, discussion techniques, student participation

Table A.2: FFT Teaching Practice Correlations and Factor Loadings

	CERR	MCP	MSB	USDT	ECL	CS	ESL	Factor 1 Loadings	Factor 2 Loadings
CERR	1							0.196	0.680
MCP	0.602***	1						0.055	0.779
MSB	0.676***	0.713***	1					-0.090	0.934
USDT	0.476***	0.413***	0.395***	1				0.790	-0.033
ECL	0.627***	0.497***	0.496***	0.569***	1			0.699	0.170
CS	0.568***	0.524***	0.464***	0.559***	0.601***	1		0.592	0.219
ESL	0.489***	0.452***	0.415***	0.627***	0.700***	0.575***	1	0.886	-0.067
UAI	0.462***	0.468***	0.416***	0.644***	0.597***	0.586***	0.667***	0.826	-0.032
Obs.					732				

Notes: First seven columns show correlations between FFT components. We use the entire sample of fourth and fifth grade teachers from both years e.g. 732 teacher-year observations. Last two columns present factor loadings from exploratory factor analysis after performing an oblique rotation of the factors, and keeping the first two factors. The first factor explains 79% of the variance in the data, and the second explains another 13%. CERR (creating an environment of respect and rapport), USDT (using questioning and discussion techniques), ECL (establishing a culture of learning), MCP (managing classroom procedures), CS (communicating with students), MSB (managing student behaviors), ESL (engaging students in learning), UAI (using assessment in instruction). See table ($A.1$) for a detailed description of each FFT variable.

Table A.3: Summary Statistics: Sample (N=2632)

	Mean	Std. Dev.	Min	Max
Grade Level	4.50	0.50	4.00	5.00
Joint Math and ELA Class	0.87	0.33	0.00	1.00
Age	9.40	0.92	7.52	12.20
Male	0.50	0.50	0.00	1.00
Gifted	0.05	0.21	0.00	1.00
Special Education	0.08	0.27	0.00	1.00
English Language Learner	0.16	0.36	0.00	1.00
White	0.25	0.43	0.00	1.00
Black	0.31	0.46	0.00	1.00
Hispanic	0.29	0.45	0.00	1.00
Asian	0.11	0.31	0.00	1.00
American Indian	0.01	0.08	0.00	1.00
Race Other	0.03	0.17	0.00	1.00
Race Missing	0.00	0.07	0.00	1.00
Math Score (Year 09-10)	0.02	0.90	-2.82	2.75
Math Score (Year 10-11)	0.04	0.90	-3.26	3.01
Percentage of Class w/ 09-10 Math Scores	0.91	0.07	0.67	1.00
Percentage of Class in Random Assignment	0.78	0.14	0.32	1.00
Teachers per Randomization Block	2.86	0.83	2.00	4.00
Randomization Block Compliance Rate	0.93	0.09	0.50	1.00

Notes: Joint Math/ELA Class refers to a self-contained course in which students learn both math and ela, the remaining courses are either math or ela only. We summarize the percentage of each class w/ prior math test scores since students new to the district will not have prior test scores. We also summarize the percentage of each class in randomization because not all students in the classes we observe were on the original randomly assigned class rosters.

Table A.4: Summary Statistics: Pre-Restricted Sample

	Mean	SD	Min	Max
Grade Level	4.52	0.50	4.00	5.00
Joint Math and ELA Class	0.85	0.36	0.00	1.00
Age	9.46	0.96	7.52	13.20
Male	0.49	0.50	0.00	1.00
Gifted	0.08	0.27	0.00	1.00
Special Education	0.09	0.29	0.00	1.00
English Language Learner	0.15	0.36	0.00	1.00
White	0.28	0.45	0.00	1.00
Black	0.34	0.48	0.00	1.00
Hispanic	0.27	0.45	0.00	1.00
Asian	0.07	0.26	0.00	1.00
American Indian	0.00	0.07	0.00	1.00
Race Other	0.02	0.15	0.00	1.00
Race Missing	0.01	0.11	0.00	1.00
Math Score (Year 09-10)	0.11	0.93	-3.14	2.84
Math Score (Year 10-11)	0.14	0.93	-3.26	3.02
Percentage of Class w/ 09-10 Math Scores	0.91	0.07	0.63	1.00
Percentage of Class in Random Assignment	0.76	0.19	0.03	1.00
Teachers per Randomization Block	3.03	1.49	1.00	12.00
Randomization Block Compliance rate	0.66	0.40	0.00	1.00
Observations		5730		

Notes: This sample corresponds to all students in the 2010-11 school year in either a fourth or fifth grade Math or Joint Math/ELA course. Since our estimation strategy leverages the random assignment of classrooms to teachers, we restrict the sample to students with a randomly assigned teacher. No further restrictions are made. Not all cells have the same number of observations.

Table A.5: Balance Tests

	Classroom Management Random Teacher	Challenge/ Student Centered Random Teacher
	(1)	(2)
Avg. Math_{t-1}	-0.022	-0.015
	(0.079)	(0.119)
IQR Math_{t-1}	0.036	0.041
	(0.107)	(0.103)
Avg Math_{t-1} Rand	-0.089	-0.046
	(0.103)	(0.124)
IQR Math_{t-1} Rand	-0.023	0.032
	(0.087)	(0.084)

Notes: Each cell corresponds to a separate regression of the dependent variable indicated in the column header on the row-variable, controlling for randomization block fixed-effects. Standard errors are clustered at the randomization block level. Columns (1) and (2) refers to a student's randomly assigned teacher's practice measured in $t-1$ (i.e., P_{rt-1} in the present notation). Rows 3 and 4 correspond to the average and IQR of math of the randomly assigned peers.

Table A.6: Balance Tests

	Classroom Mgmt Random Teacher	Challenge/ Student Centered Random Teacher	Avg. Math of Observed Peers	IQR Math of Observed Peers	Avg Math of Assigned Peers	IQR Math of Assigned Peers
	(1)	(2)	(3)	(4)	(5)	(6)
Math_{t-1}	-0.021	-0.005	0.050	-0.029	0.053	0.006
	(0.020)	(0.024)	(0.048)	(0.036)	(0.048)	(0.025)
ELL	-0.048	-0.015	-0.200	0.025	-0.197	-0.023
	(0.059)	(0.061)	(0.129)	(0.124)	(0.136)	(0.091)
Gifted	-0.033	-0.053	0.491**	0.160	0.274	0.230*
	(0.075)	(0.144)	(0.227)	(0.107)	(0.175)	(0.123)
Special Educ.	0.118**	0.089	-0.128*	0.043	-0.055	0.028
	(0.059)	(0.057)	(0.065)	(0.084)	(0.055)	(0.066)
Male	0.008	0.002	-0.023	-0.008	-0.035	-0.025
	(0.013)	(0.015)	(0.019)	(0.015)	(0.023)	(0.018)
White	0.011	-0.044	0.035	0.011	-0.039*	-0.014
	(0.029)	(0.032)	(0.042)	(0.036)	(0.023)	(0.032)
Black	0.005	0.001	0.005	0.041	0.055**	0.044
	(0.028)	(0.031)	(0.046)	(0.048)	(0.026)	(0.048)
Hispanic	-0.059**	-0.046	-0.036	-0.029	-0.022	-0.017
	(0.028)	(0.034)	(0.029)	(0.043)	(0.028)	(0.046)
Asian	0.087	0.142**	0.070*	-0.037	0.053	0.000
	(0.054)	(0.055)	(0.039)	(0.066)	(0.044)	(0.064)
American Indian	0.062	0.176	-0.339	0.247	-0.215	0.076
	(0.145)	(0.118)	(0.205)	(0.174)	(0.212)	(0.139)
Race Other	0.070	0.063	-0.083	-0.058	-0.074**	-0.044
	(0.066)	(0.088)	(0.050)	(0.048)	(0.037)	(0.053)

Notes: Each cell corresponds to a separate regression of the dependent variable indicated in the column header on the row-variable, controlling for randomization block fixed-effects. Standard errors are clustered at the randomization block level. Columns (1) and (2) refers to a student's randomly assigned teacher's practice measured in $t-1$ (i.e., P_{rt-1} in the present notation). Columns (3) and (4) use the actual classroom composition whereas columns (5) and (6) focus on the peers who were initially assigned to be grouped with the student. Row 1 of Columns (3) to (6) control for the average peer prior achievement of the randomization block, to deal with the mechanical negative correlation highlighted in Guryan et al. (2009).

Table A.7: Effects of Teaching Practice without Classroom Interactions

	Actual Teacher		Random Teacher		IV Actual with Rand. Teacher	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Classroom Mgmt	0.005	0.005	0.009	0.009	0.010	0.009
	(0.025)	(0.024)	(0.022)	(0.021)	(0.025)	(0.024)
C.M. × Math_{t-1}		0.018		0.022*		0.023*
		(0.013)		(0.013)		(0.013)
P-value (joint signif. of teaching practice)		0.380		0.239		0.219
F-Stat. (first stage)†					251.3	167.1
Panel B						
Challenge/ Student-Centered	0.021	0.019	0.025	0.023	0.029	0.026
	(0.022)	(0.022)	(0.021)	(0.021)	(0.024)	(0.024)
C.S.C. × Math_{t-1}		0.017		0.025*		0.026*
		(0.013)		(0.013)		(0.014)
P-value (joint signif. of teaching practice)		0.195		0.042		0.029
F-Stat. (first stage)†					279.8	186.6

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Panel A and B correspond to different regressions with math as the dependent variable. Lagged teaching practices are used and sample size is 2632. These regressions include randomization block fixed effects, levels and squared-terms in prior math and average peer prior math, as well as CKT and student characteristics listed in Table A.3. † Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test.

Table A.8: Comparison between the Hausman Estimator and ITT-IV specifications

	MCP ITT	MCP IV	FFT MCP-MSB-CERR	ESL ITT	ESL IV	FFT ESL-USDT
	(1)	(2)	(3)	(4)	(5)	(6)
Teaching Practice	0.004	0.001	0.011	0.028	0.022	0.021
	(0.019)	(0.021)	(0.027)	(0.019)	(0.020)	(0.031)
T.P. × Math_{t-1}	0.009	0.014	0.011	0.001	0.014	0.013
	(0.014)	(0.015)	(0.022)	(0.012)	(0.013)	(0.021)
T.P. × Peer Math	0.052***	0.105***	0.111***	0.005	0.043**	0.019
	(0.019)	(0.033)	(0.039)	(0.014)	(0.017)	(0.044)
T.P. × IQR Math	−0.035**	−0.004	−0.008	−0.047***	−0.055***	−0.063**
	(0.016)	(0.025)	(0.033)	(0.016)	(0.014)	(0.029)
P-value joint signif. T.P.	0.000	0.000		0.038	0.000	
First Stage F-Stat.†		21.7			12.5	
Hansen J P-value††		0.522			0.643	
p2 load			1.080			0.804
p3 load			0.859			0.837

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Sample size is 2632. Managing student behaviors (MSB), Managing classroom procedures (MCP), Creating an environment of respect and rapport (CERR), Engaging students in learning (ESL), Using questioning and discussion techniques (USDT). The ITT columns uses randomly assigned MCP or ESL scores as "Practice." The IV columns use all other practices that load on classroom management to instrument for MCP, and likewise for ESL with challenge/student-centered practices. Practices are for the randomly assigned teacher measured at $t-1$. We use efficient GMM estimator and FFT MCP-MSB-CERR uses our adapted Hausman estimator to correct for measurement error, where MCP is the anchor, and MSB is used to construct moment conditions. FFT ESL-USDT is similar but uses the average of all other challenge/student-centered practices as the third measurement since we are overidentified. The specification is identical to that in Table (2) except here we do not include controls for student characteristics. † Reports the Kleibergen-Paap rk Wald statistic. †† Reports p-value from Hansen's J statistic test of overidentifying restrictions."p2 load" and "p3 load" are the recovered measurement parameters described in Appendix D.2. Standard errors are clustered at the randomization block level, and with the adapted Hausman estimator we bootstrap standard errors with 200 repetitions.

Table A.9: Contemporaneous Teaching Practice and Classroom Composition

	ITT		IV	ITT	IV	
	Time $t-1$		Time t	Time $t-1$	Time t	
	Practice		Practice	Practice	Practice	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Classroom Management	0.008			0.010	0.049*	0.040
	(0.019)			(0.018)	(0.026)	(0.058)
C.M. $\times$ Math$_{t-1}$	0.011			0.012	0.004	0.026
	(0.012)			(0.012)	(0.018)	(0.021)
C.M. $\times$ Avg. Peer Math$_{t-1}$	0.07***	0.085***	0.218***	0.082***	0.213***	0.209***
	(0.025)	(0.022)	(0.065)	(0.021)	(0.062)	(0.061)
C.M. $\times$ IQR Peer Math$_{t-1}$	-0.017					
	(0.019)					
First Stage F-Stat.†			28.400		31.563	4.191
Panel B						
Challenge/Student-Centered	0.018			0.022	0.072***	-0.006
	(0.020)			(0.019)	(0.026)	(0.127)
C.S.C $\times$ Math$_{t-1}$	0.012			0.018	0.008	0.041
	(0.012)			(0.012)	(0.014)	(0.033)
C.S.C $\times$ Avg Peer Math$_{t-1}$	0.031**					
	(0.014)					
C.S.C. $\times$ IQR Peer Math$_{t-1}$	-0.053^{***}	-0.063^{***}	-0.178^{***}	-0.060^{***}	-0.181^{***}	-0.174^{***}
	(0.014)	(0.016)	(0.052)	(0.014)	(0.051)	(0.051)
First Stage F-Statistic†			26.896		24.538	2.107

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Sample size is 2632. Randomly assigned teachers are used throughout. Panel A and B correspond to different regressions with math as the dependent variable. These regressions include randomization block fixed effects and controls for the level and a squared term of prior math achievement and average peer prior achievement, IQR in peer prior achievement, along with the peer variables squared and interactions with each other and lagged math achievement. Controls for CKT and student characteristics listed in Table A.3 also included. † Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test.

Table A.10: Individual FFT Subdomain Regressions

Panel A	Creating environment of respect & rapport	Managing classroom procedures	Managing student behaviors	Establish culture of learning
Practice	0.020	0.004	0.003	0.010
	(0.018)	(0.019)	(0.017)	(0.022)
Practice × Math_{t-1}	0.011	0.009	0.009	0.006
	(0.012)	(0.013)	(0.013)	(0.011)
Practice ×	0.057***	0.052***	0.072***	0.049**
Avg Peer Math_{t-1}	(0.020)	(0.019)	(0.026)	(0.019)
Practice ×	-0.019	−0.035**	-0.012	−0.040***
IQR Peer Math_{t-1}	(0.017)	(0.016)	(0.019)	(0.015)
P-value (joint signif. of teaching practice)	0.000	0.000	0.002	0.000

Panel B	Engaging students in learning	Using questioning and discussion	Using assessment in instruction	Communicating with students
Practice	0.028	0.011	0.016	0.014
	(0.018)	(0.017)	(0.021)	(0.018)
Practice × Math_{t-1}	0.001	0.011	0.020*	0.015
	(0.012)	(0.014)	(0.011)	(0.012)
Practice ×	0.005	0.020*	0.037**	0.034**
Avg Peer Math_{t-1}	(0.014)	(0.012)	(0.016)	(0.014)
Practice ×	−0.047***	−0.046***	−0.039***	−0.070***
IQR Peer Math_{t-1}	(0.015)	(0.017)	(0.014)	(0.026)
P-value (joint signif. of teaching practice)	0.035	0.017	0.000	0.000

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Panel A and B correspond to different regressions with math as the dependent variable. Lagged teaching practices are used and sample size is 2632. These regressions include randomization block fixed effects and controls for the level and a squared term of prior math achievement, average peer prior achievement, IQR of peer prior achievement as well as CKT and student characteristics listed in Table A.3. The first 3 subdomains correspond to classroom management, the remainder to challenge/student-centered.

Table A.11: Teaching Practices and Alternative Teacher "Quality" Controls: Full Results

	Random Teacher	IV Actual with Random Teacher	Random Teacher Alt. Teacher Control: CKT	7C	PSVY
	(1)	(2)	(3)	(4)	(5)
Classroom Management	-0.012	-0.016	-0.014	-0.016	-0.015
	(0.020)	(0.022)	(0.020)	(0.020)	(0.019)
C.M. × Math_{t-1}	0.004	0.004	0.011	0.004	0.003
	(0.020)	(0.021)	(0.019)	(0.019)	(0.019)
C.M. × Peer Math	0.076**	0.087**	0.077**	0.076**	0.076***
	(0.029)	(0.036)	(0.030)	(0.029)	(0.027)
C.M. × IQR Math	0.026	0.035	0.026	0.026	0.026
	(0.022)	(0.026)	(0.022)	(0.023)	(0.021)
C.M.	0.026	0.025	0.026	0.026	0.011
	(0.023)	(0.025)	(0.022)	(0.022)	(0.024)
C.M. × Math_{t-1}	0.010	0.011	0.002	0.016	0.005
	(0.020)	(0.021)	(0.020)	(0.019)	(0.019)
C.M. × Peer Math	-0.010	-0.009	-0.010	-0.010	-0.005
	(0.019)	(0.022)	(0.019)	(0.019)	(0.019)
C.M. × IQR Math	-0.062***	-0.071***	-0.063***	-0.057**	-0.054**
	(0.017)	(0.019)	(0.017)	(0.021)	(0.021)
CKT	-0.007	-0.011	-0.008	-0.006	-0.013
	(0.016)	(0.019)	(0.016)	(0.016)	(0.018)
Alt. Teacher Control				-0.006	0.055***
				(0.019)	(0.017)
T.C. × Math_{t-1}			0.044***	-0.029**	0.032**
			(0.014)	(0.013)	(0.013)

T.C. × Peer Math			-0.019	-0.007	-0.016
			(0.018)	(0.020)	(0.016)
T.C. × IQR Math			-0.012	-0.017	-0.003
			(0.021)	(0.021)	(0.016)
T.C. missing					-0.591***
					(0.142)
T.C. missing × Math_{t-1}					0.025
					(0.046)
T.C. missing × Peer Math					0.060
					(0.045)
T.C. missing × IQR Math					0.015
					(0.055)
Math_{t-1}	0.724***	0.725***	0.723***	0.723***	0.722***
	(0.017)	(0.017)	(0.016)	(0.016)	(0.018)
Math^2_{t-1}	-0.043***	-0.043***	-0.044***	-0.042***	-0.045***
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Peer Math × Math_{t-1}	-0.000	-0.001	-0.001	-0.003	0.002
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
IQR Math × Math_{t-1}	0.034**	0.035***	0.040***	0.031**	0.044***
	(0.013)	(0.013)	(0.013)	(0.013)	(0.012)
Peer Math × IQR Math	-0.052***	-0.053***	-0.057***	-0.053***	-0.043**
	(0.016)	(0.016)	(0.017)	(0.016)	(0.017)
Peer Math × IQR Math × Math_{t-1}	-0.021	-0.021	-0.019	-0.023	-0.016
	(0.014)	(0.014)	(0.015)	(0.014)	(0.014)

Peer Math	-0.008	-0.008	-0.009	-0.007	-0.012
	(0.026)	(0.026)	(0.025)	(0.026)	(0.027)
Peer Math^2	-0.010	-0.013	-0.009	-0.009	-0.014
	(0.014)	(0.013)	(0.014)	(0.014)	(0.016)
IQR Math	-0.015	-0.018	-0.017	-0.019	-0.008
	(0.023)	(0.024)	(0.022)	(0.024)	(0.026)
IQR Math^2	-0.008	-0.009	-0.009	-0.010	-0.002
	(0.013)	(0.014)	(0.013)	(0.014)	(0.014)
ELL	0.008	0.015	0.011	0.007	0.009
	(0.038)	(0.039)	(0.039)	(0.038)	(0.038)
Gifted	0.195***	0.188***	0.195***	0.192***	0.198***
	(0.055)	(0.054)	(0.054)	(0.057)	(0.056)
Male	-0.001	-0.004	-0.002	-0.001	-0.001
	(0.021)	(0.021)	(0.021)	(0.021)	(0.021)
Special Educ.	-0.111**	-0.110**	-0.110**	-0.112**	-0.108**
	(0.043)	(0.043)	(0.042)	(0.044)	(0.043)
Black	-0.157***	-0.159***	-0.148***	-0.154***	-0.156***
	(0.033)	(0.032)	(0.033)	(0.033)	(0.034)
Hispanic	-0.047	-0.051	-0.044	-0.046	-0.049
	(0.035)	(0.034)	(0.035)	(0.036)	(0.035)
Asian	0.076**	0.069*	0.082**	0.078**	0.070*
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
American Indian	-0.045	-0.050	-0.036	-0.045	-0.048
	(0.108)	(0.107)	(0.107)	(0.109)	(0.111)
Race Other	0.013	0.013	0.016	0.016	0.013
	(0.047)	(0.046)	(0.047)	(0.047)	(0.047)
Race Missing	-0.040	-0.046	-0.012	-0.041	-0.044
	(0.069)	(0.066)	(0.073)	(0.067)	(0.061)
R^2	0.649	0.708	0.651	0.650	0.652
P-value joint signif of C.M. & C.S.C.	0.000	0.000	0.000	0.000	0.002

P-value joint signif of T.C.		0.052	0.172	0.013
First Stage F-Statistic†	27.717			

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Standard errors are clustered at the randomization block level. Sample size is 2632. Dependent variable is math and teaching practices are measured at $t-1$. Regressions include randomization block fixed effects and controls for the level and a squared term of prior math achievement and average and IQR of peer prior achievement, their square and all pairwise interactions of peer variables and prior achievement, as well as student characteristics listed in Table A.3. Even columns also include the IQR in peer prior achievement. † Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test. CKT denotes Content Knowledge for Teaching assessment, $7C$ denotes overall student survey teacher ratings based on Tripod and $PSVY$ denotes principal assessments of teacher quality.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1573	Stephan Heblich Stephen J. Redding Daniel M. Sturm	The Making of the Modern Metropolis: Evidence from London
1572	Johannes Beohm Ezra Oberfield	Misallocation in the Market for Inputs: Enforcement and the Organization of Production
1571	Hans R. A. Koster Edward W. Pinchbeck	How do Households Value the Future? Evidence from Property Taxes
1570	Holger Breinlich Elsa Leromain Dennis Novy Thomas Sampson Ahmed Usman	The Economic Effects of Brexit – Evidence From the Stock Market
1569	J. Vernon Henderson Sebastian Kriticos Dzhamilya Nigmatulina	Measuring Urban Economic Density
1568	Philippe Bracke Silvana Tenreyro	History Dependence in the Housing Market
1567	Ester Faia Sebastien Laffitte Gianmarco Ottaviano	Foreign Expansion, Competition and Bank Risk
1566	Brian Bell Rui Costa Stephen Machin	Why Does Education Reduce Crime?
1565	Richard Murphy Felix Weinhardt Gill Wyness	Who Teaches the Teachers? A RCT of Peer-to-Peer Observation and Feedback in 181 Schools

1564	Paola Conconi Giovanni Facchini Max F. Steinhardt Maurizio Zanardi	The Political Economy of Trade and Migration: Evidence from the U.S. Congress
1563	Stephen Gibbons Stephan Heblich Ted Pinchbeck	The Spatial Impacts of a Massive Rail Disinvestment Program: The Beeching Axe
1562	Helen Johnson Sandra McNally Heather Rolfe Jenifer Ruiz-Valenzuela, Robert Savage Janet Vousden Clare Wood	Teaching Assistants, Computers and Classroom Management: Evidence from a Randomised Control Trial
1561	Daniel J. Graham Stephen Gibbons	Quantifying Wide Economic Impacts of Agglomeration for Transport Appraisal: Existing Evidence and Future Directions
1560	Daniel S. Jacks Dennis Novy	Market Potential and Global Growth over the Long Twentieth Century
1559	Alan Manning Paolo Masella	Diffusion of Social Values through the Lens of US Newspapers
1558	Jan David Bakker Stephan Maurer Jörn-Steffen Pischke Ferdinand Rauch	Of Mice and Merchants: Trade and Growth in the Iron Age
1557	Giuseppe Berlingieri Sara Calligaris Chiara Criscuolo	The Productivity-Wage Premium: Does Size Still Matter in a Service Economy?
1556	Christian A.L. Hilber Olivier Schöni	The Economic Impacts of Constraining Second Home Investments

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE