Teacher Effectiveness and Classroom Composition*

Esteban M. Aucejo† Patrick Coate‡ Jane Cooley Fruehwirth§

Arizona State University *NCCI* *University of North Carolina*

Sean Kelly¶ Zachary Mozenter‖

University of Pittsburgh *University of North Carolina*

February 20, 2018

Preliminary and Incomplete

Abstract

This paper bridges the gap between the teacher effectiveness and peer effects literatures, by studying how the effectiveness of different teaching practices vary with the classroom composition. We combine random assignment of teachers to classrooms with rich measures of teaching practice based on trained observers using a popular teacher evaluation protocol. We find that good classroom behavior management skills create an environment where students benefit more from peer average initial achievement. On the other hand, student-centered practices are most effective when there is low variation in the initial achievement of classmates, but do not vary notably with average initial achievement. This has important implications for measuring teacher effectiveness and peer effects, and for guiding teaching practice in different classroom contexts.

Keywords: Teacher, Practices, Peer Effects, Effectiveness

JEL Classification Codes: I2, I20, I21

*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 of the U.S. Department of Education. Fruehwirth also thanks the British Academy and the Leverhulme Trust's Philip Leverhulme Prize. We also thank Pat Bayer, Joe Hotz, Ju Hyun Kim, Doug Staiger, Valentin Verdier for helpful comments and conversations.

†Department of Economics, W.P. Carey School of Business, Arizona State University & CEP. Email: Esteban.Aucejo@asu.edu

‡National Council on Compensation Insurance. Email: pecoate@gmail.com

§Department of Economics and Carolina Population Center, University of North Carolina, Chapel Hill. Email: jane_fruehwirth@unc.edu

¶School of Education, University of Pittsburgh. Email: spkelly@pitt.edu

‖Department of Economics, University of North Carolina, Chapel Hill. Email: zmozent@gmail.com

1 Introduction

The Coleman Report of 1966 highlighted the importance of both teachers and peers in improving academic achievement. Moreover, it spawned a large literature and policy debate regarding teacher and peer influence, particularly in understanding racial achievement gaps. Teachers skills, effort and/or pedagogical practices are critical to understanding disparities in rates of learning across students (Gamoran et al., 2000; Rivkin et al., 2005). Similarly, there are a large number of channels through which peers are thought to affect educational outcomes. For example, students could benefit from interactions with higher ability peers and suffer from the presence of disruptive classmates (Sacerdote, 2011; Epple and Romano, 2010). Though, these literatures are vast, they have largely treated teachers and peers in isolation.

Treating teachers and peers as separable influences on learning has (at least) two important limitations. First, it fails to acknowledge that the effectiveness of different teachers/teaching practices could depend on the characteristics of the classroom.[1] For example, the benefits of student-centered teaching practices may vary depending on the heterogeneity in initial achievement of a student's classmates. Second, it fails to acknowledge the fundamental role teachers can play in determining the nature of the interactions among classmates. For instance, peer effects could be amplified by teaching practices that promote a learning dialogue among students. In a similar vein, disruptive behavior is an endogenous response to the teacher's skills in managing classroom behaviors and engaging students. Thus, the negative consequences of disruptive classmates could be largely preempted by effective teachers. We seek to fill this gap in the literature, exploring complementarities between teachers and classroom composition in achievement production, and will demonstrate why it is critical for understanding both teacher and peer effects.

We illustrate the pervasiveness of the potential complementarities in teaching practice and classroom composition in a simple theoretical framework. We focus on spillovers from peer initial achievement, which are the most-studied type of peer spillover in the literature. We show that even when the learning production function does not directly depend on the interaction between

[1] Teaching practices do 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. behavior management).

teaching practice and peer initial achievement, a complementarity between teachers and peers will emerge indirectly through students' endogenous responses to teaching practices. One such example is when teachers with better behavior management practices make misbehavior more costly, and students benefit more from their peers if they behave well. Based on this framework, we derive a set of equations that guide our empirical strategy.

Two important barriers have hindered an unified analysis of teacher effectiveness and peer effects. First, detailed longitudinal data on teaching practices on a large scale are relatively rare. Second, problems of selection and endogeneity (i.e. allocation of teachers into classrooms, and endogenous changes in teaching practices) have prevented the implementation of credible empirical strategies. We overcome these problems by exploiting the information collected in the Measures of Effective Teaching (MET) Longitudinal Database. The key feature of this database is that it provides rich information on teaching practices in a context where teachers were 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, *Framework for Teaching Evaluation Instrument* (Danielson, 2011).

The random assignment of teachers eliminates one of the most important confounding factors for measuring teacher effectiveness, the systematic matching of students to classrooms 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. This is easily addressed using typical strategies in the randomized control trial literature given that we observe the randomly-assigned teacher as well as the actual teacher. Second, the experimental design did not mandate random assignment of students to classrooms. That said, the random assignment of teachers to classrooms is enough to obtain consistent results of the complementarities between teaching practice and classroom composition, applying results developed in Bun and Harrison (2014) and Nizalova and Murtazashvili (2014).[2] Third, if teachers choose practices to maximize student achievement, the observed teaching practice could be endogenous to the class-

[2]That said, this may limit our ability to infer the overall effect of peers, depending on whether classroom composition is non-random.

room composition. We address this primarily by exploiting the availability of prior year teaching practices, thus capturing more a teachers' proclivity toward certain practices. Fourth, teaching practice is measured with error. We exploit multiple measures of teaching practice through factor models to identify what aspects are separable in the data. While we rely primarily on averages of multiple measures of teaching practices to address measurement error, we show robustness of a number of other approaches. This is complicated in our setting because of nonlinearities in the effect of teaching practice inherent in our equations. We adapt the estimation approach developed by Hausman et al. (1991) for nonlinear error in variables models to apply to our case, a panel model where the nonlinearity takes the form of complementarities.

We make several important contributions to the literature. First, we demonstrate how failing to capture the heterogeneity in the effectiveness of teaching practice by classroom composition leads us to understate the importance of measured teaching practices and even, in some cases, to infer that the practice does not matter when in fact the effects are sizable in certain classrooms. This will provide insight into why observable teacher measures generally do a poor job of capturing teacher quality (e.g. Rivkin et al., 2005). From a policy perspective then, understanding this type of heterogeneity is crucial for identifying what teaching practices matter in what classroom contexts.

Second, we demonstrate that failure to allow for complementarities with teaching practice may severely understate the benefits of peers. To the extent that teaching practices vary widely across settings, this could help to reconcile significant disparities in the estimated importance of peers across a number of studies (Sacerdote, 2011). More importantly, it opens new avenues of constructively improving achievement through teaching that makes best use of a given classroom composition.

This work connects closely to a number of recent studies that consider heterogeneity in teacher effectiveness by student background characteristics (Lavy, 2015; Araujo et al., 2014; Fox, 2016; Konstantopoulos, 2009). For instance, Lavy (2015) finds larger effects of student-centered teaching for girls and low-SES students. Connor et al. (2004) show larger effects of some types of student-centered practices for children with higher initial achievement. Finally, Konstantopoulos (2009) finds somewhat larger effects of teacher effectiveness for high-SES students. 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 initial achievement.

Our study also provides useful complementary evidence to the value-added literature which argues fairly persuasively that teachers matter, but are not able to identify basic teacher characteristics that are consistently associated with teacher effectiveness (Rivkin et al., 2005; Chetty et al., 2014; Rothstein, 2010). A number of other studies have used the MET data to identify effective teachers. Already studies from the MET project have generated important insights (Cantrell and Kane, 2013). 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) also show that the different metrics of teacher effectiveness (value-added, classroom observation video scores and student survey reports) have important commonalities. In the present study, we shift the emphasis from identifying effective teachers to analyzing the elements of teaching that are, on average, most effective for different kinds of classrooms.

Peer effect studies have focused more squarely on how peer effects vary by student background characteristics because of the important implications of this type of heterogeneity to tracking and desegregation policies. The evidence is somewhat mixed, with some studies showing a great deal of heterogeneity in peer effects by race and initial achievement (For instance, see Burke and Sass, 2006; Fruehwirth, 2013; Gibbons and Telhaj, 2006; Hanushek et al., 2009; Hanushek and Rivkin, 2009; Hoxby and Weingarth, 2005; Lavy et al., 2012, among others). None of these consider heterogeneity by teaching practice which may, similarly to the case above, be an important confounder.

Our main findings indicate that disregarding complementarities between classroom characteristics and teaching practices make it difficult to detect the value of teaching practices. We show that student-centered practices are more effective when classrooms have less heterogeneity in initial achievement. In addition, we show that teacher practices aimed at managing behavior are more effective when classroom have higher average initial achievement. This highlights the intuition that

students cannot benefit from better-prepared peers if they are not behaving well in class. This is consistent with the understanding that behavior management is even an important challenge in higher-achieving classrooms, though the sources of disengagement and misbehavior may be different from in lower-achieving classrooms (Shernoff et al., 2003). We find on the flipside that peer effects become more important as well when teaching practice is taken into account. We show robustness of these findings to the variety of concerns discussed above.

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. Section 4 discusses our empirical strategy. Section 5 presents our main findings, followed by the conclusion in Section 6.

2 Data

The Measures of Effective Teaching (MET) Longitudinal Database provides detailed information on teaching practices, student outcomes, and classroom composition from six large urban public school districts in the United States over two academic years (2009-2010 and 2010-2011).[3] The data are linked to district administrative records, giving us access to detailed student information, most important, 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, and teacher background characteristics (e.g. sex, race/ethnicity, degree status, years of teaching experience in the current school, and measures related to teacher aptitude, such as, the Content Knowledge for Teaching (CKT) assessment, and school principal evaluations).[4] Finally, a key aspect of the MET data is that teachers from more than 200 schools were randomly assigned

[3]These 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. Kane and Staiger (2012) provides a detailed description on 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.

[4]The purpose of the CKT is to assess whether a prospective elementary teacher has the content knowledge needed at the time of entry to the profession in the areas of reading and language arts, mathematics, science, and social studies. It is designed for teacher candidates seeking a generalist elementary school license.

within school and grade to classrooms of students during the second academic year (2010-2011).[5] This random allocation of teachers into classrooms plays a key role in our empirical strategy.

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.[6]

2.1 Measuring Teaching Practice

We make use of a well-known general classroom observation protocol that measures teaching practices, i.e. Framework for Teaching (FFT). Protocols like FFT are becoming increasingly important from a policy perspective, because a number of school districts have begun to use them for evaluation purposes (AIR, 2013). FFT is a research-based protocol developed by educational experts to assess teacher effectiveness across subjects. According to MET project (2010b), "*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).*"[7] The protocol divides teaching components into four domains, with the MET database rating teachers on two of them: classroom environment and instruction. We observe scores for eight different components of these two domains by a median of seven different highly trained, independent raters, many of them current or former teachers.[8] 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

[5]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".

[6]Appendix A provides a detailed description of the sample selection.

[7]Of the MET observation protocol, 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 teacher practices.

[8]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 7 provides a description of each of the sub-components of the FFT protocol.

observational data and help us to address measurement error, as we discuss further in Section 4.

Though FFT was designed so that each component 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 8 shows the correlations between the different components and the loadings on each FFT component after performing an oblique rotation of the factors.[9] 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 *student-centered practices* that encourage classroom dialogue and student involvement.[10] The sub-scales that load on the second factor are creating an environment of respect and rapport, managing student behaviors and managing classroom procedures, all of which capture domains of instruction related to *behavior management*. Taken together the factors explain 92% of the total variance in the data.[11] Finally, it is important to emphasize that these groupings of sub-scales are guided both by the pedagogical theories underlying the construction of these protocols and what information is separable from these measures. Our empirical strategy will mainly make use of averages across the sub-scales that according to the exploratory factor analysis correspond to each broad practice (i.e. behavior management and student-centered practices), but we also explore other ways of addressing measurement error, as described in detail in Section 4.[12]

[9]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. They are also similar if we extract rater fixed effects and video quality prior to performing the factor analysis. Orthogonal rotations also provide a similar conclusion.

[10]We have chosen the term "student-centered practices" to try to capture the overall emphasis of the model item. 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).)

[11]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.

[12]We also replicated our empirical strategy using both confirmatory factor analysis and principal component as alternative measures of student-centered and behavior management practices. Results in all cases are similar.

2.2 Summary Statistics

Table 1 reports summary statistics (age, race, gender, proportion in gifted classes and special education, proportion that is free and reduced price lunch eligible, and English language learners) corresponding to the students in our final sample.[13] For example, around 8% of them required special education and 64% are free school lunch eligible. This is a racially-diverse sample; 31% of students are black, 24% are white, 32% are Hispanic, and 9% are Asian, indicating that the school districts included in our data are not necessarily representative of the whole US population of students. The bottom part of Table 1 further characterizes the data by displaying the number of districts (5), schools (45), teachers (183), and randomization blocks (70) in our final sample.

In terms of classroom and teacher characteristics, Table 2 displays summary statistics corresponding to the the FFT domains and classroom prior achievement average and inter-quantile range. Raw standard deviations reported in Column (2) show substantial variation in these variables, however, given that randomization of teachers into classrooms was performed within school-grade level (i.e. randomization block), then it is important to assess whether there is sufficient variation within these blocks. In this regard, we report in the last two columns of Table 2 standard deviations within and between randomization blocks. We find considerable within-randomization block variation in teaching practice that is only marginally smaller than that between blocks. Similarly, we also find important variation in peer characteristics within blocks.

To conclude, we analyze whether the random assignment of teachers into classrooms (within randomization block) has been successful by implementing balancing tests. First, we check that teaching practices at $t-1$ of randomly assigned teachers do not correlate with observed classroom composition at t and with students' characteristics. Second, we analyze whether these tests look different using initially assigned classroom composition instead. Appendix Table 10 shows regressions of randomly-assigned teaching practices (behavior management, and student-centered practices) on IQR of peer prior achievement of the actual and initially-assigned classroom composition.[14] This analysis shows that the randomization performed by the MET project is reliable given that almost

[13] Appendix A describes in detail the different restrictions that we imposed to the original sample in order to obtain our final sample of 3322 students. Appendix Table 9 shows summary statistics of the full sample.

[14] Each cell corresponds to a separate regression.

Table 1: Summary Statistics: Restricted Sample (N=3322)

	Mean	Std. Dev.	Min	Max
Grade Level	4.492	0.50	4.00	5.00
Joint Math/ELA Class	0.872	0.33	0.00	1.00
Age	9.416	0.97	7.52	12.40
Male	0.495	0.50	0.00	1.00
Gifted	0.045	0.21	0.00	1.00
Special Education	0.082	0.27	0.00	1.00
English Language Learner	0.179	0.38	0.00	1.00
White	0.242	0.43	0.00	1.00
Black	0.312	0.46	0.00	1.00
Hispanic	0.315	0.46	0.00	1.00
Asian	0.091	0.29	0.00	1.00
American Indian	0.006	0.08	0.00	1.00
Race Other	0.028	0.16	0.00	1.00
Math Score (Year 09-10)	0.019	0.89	-2.82	2.75
Math Score (Year 10-11)	0.046	0.90	-3.26	3.02
Unique Districts	5	-	-	-
Unique Classes	183	-	-	-
Unique Schools	45	-	-	-
Unique Randomization Blocks	70	-	-	-
Unique Teachers	183	-	-	-
Percentage of Class w/ 09-10 Math Scores	0.915	0.07	0.67	1.00
Percentage of Class in Random Assignment	0.789	0.14	0.32	1.00
Teachers per Randomization Block	2.896	0.82	2.00	4.00
Randomization Block Compliance Rate	0.931	0.09	0.50	1.00

Notes: See Appendix A for a description of how this sample was obtained. 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 2: 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	-2.27	2.96	0.86	0.54
IQR Peer Math$_{t-1}$	0	1	-2.42	3.01	0.78	0.69
Avg Peer Math$_{t-1}$ (random)	0	1	-2.69	2.98	0.86	0.54
IQR Peer Math$_{t-1}$ (random)	0	1	-2.32	4.3	0.78	0.7
Teaching Practices						
Student Centered	0	1	-3.05	2.24	0.75	0.7
Behavior Management	0	1	-3.18	2.27	0.74	0.64
FFT Domains						
CERR	2.79	0.34	1.67	3.5	0.24	0.23
USDT	2.21	0.36	1.25	3.25	0.27	0.26
ECL	2.61	0.34	1.67	3.5	0.26	0.23
MCP	2.74	0.37	1.67	3.5	0.27	0.25
CS	2.68	0.33	2.00	3.33	0.24	0.24
MSB	2.81	0.35	1.67	3.5	0.25	0.24
ESL	2.54	0.35	1.67	3.5	0.23	0.27
UAI	2.42	0.37	1.33	3.5	0.27	0.26

Notes: The sample size is 3322 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 subcomponents of FFT are 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), and UAI (using assessment in instruction), as described further in Table (7). *Behavior management* is the standardized average of MSB, MCP and CERR. *Student-centered* is the standardized average of ECL, CS, ESL, UAI, and USDT. The last two columns decompose the standard deviation for each variable into between randomization block and within randomization block components.

all coefficients are not statistically significantly different from 0. In addition, these results show that any reallocation of students across classrooms after the initial random assignment does not seem to lead to statistically significant correlations of our classroom composition with the teaching practice.[15]

3 Model

Standard models of educational achievement treat teachers and peers as separable inputs. Our premise is that this may at best understate and at worst lead to misleading conclusions about how teachers and peers shape achievement production. We motivate here 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 an environment where peers matter more for each student's achievement, creating more of a team production climate. Each student's questions or contributions to a discussion could, in theory, have either a positive or negative externality on peer achievement. Alternatively, it is also possible that the interactions arise through a model where the teaching practice affects achievement indirectly through students' behavioral responses to the practice. We hypothesize that classroom behavior management practices often fall into this latter category. While the production technology channel is straightforward, the latter needs further motivation.

Let Y_{it} denote achievement of a student i at time t. Let the index $c = c(i,t)$ denote i's classroom in period t and then the vector of classroom peer achievement excluding i is denoted $Y_{-ict} = (Y_{1t}, ..., Y_{i-1,t}, Y_{i+1,t}, ..., Y_{Nt})$. A student's class is also assigned to 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, some moment of the prior achievement distribution of their time t classmates ($m(Y_{-ict-1})$). The less standard input to production that we introduce is student behavior, b_{it}, which we take to be unidimensional for simplicity. We conceptualize behavior

[15] Furthermore, in section 4 we also show that endogeneity of classroom composition would not bias estimates of the interaction of teaching practice with classroom composition, as long as the teaching practice is randomly assigned, which these results support.

broadly as some aggregate of student attentiveness, engagement and/or effort.

$$Y_{it} = \beta_0 + \beta_b b_{it} + \beta_{by} b_{it} Y_{it-1} + \beta_{b\bar{y}} b_{it} m(Y_{-ict-1}) + \beta_y Y_{it-1} + \beta_{\bar{y}} m(Y_{-ict-1}) + \\ + \beta_p P_j' + \beta_{py} P_j' Y_{it-1} + \beta_{p\bar{y}} P_j' m(Y_{-ict-1}) + \epsilon_{it}, \quad (1)$$

where ϵ_{it} denotes the residual. This specification permits that the marginal value of behavior is increasing in the child's own initial achievement. Furthermore, the marginal benefits of behavior vary with the composition of the classroom. For instance, if $m(Y_{-ict-1})$ is average peer initial achievement, this would allow that the returns to good behavior are higher in a classroom where peers are higher-achieving.

Students choose their behavior to maximize their expected utility from achievement. There is a cost to behavior and the cost of bad behavior (or modeled instead as the benefit of good behavior) is increasing in the teaching practice (P_j), 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 simply

$$b_{it}^* = \frac{\gamma_y}{\gamma_b}(\beta_b + \beta_{by} Y_{it-1} + \beta_{b\bar{y}} m(Y_{-ict-1})) + \frac{\gamma_{bp}}{\gamma_b} P_j'.$$

Behavior is increasing in initial achievement, peer initial achievement and teaching practice. This further permits that both behavior management practices and student-centered practices can affect behavior directly, depending on the values of γ_{bp}. We expect that this channel is more relevant for behavior management, but in principle by engaging students more through student-centered practices, teachers could also affect the value of behavior.

Plugging utility-maximizing behavior into the achievement production function, we have the

reduced form

$$
\begin{aligned}
Y_{it}^{*} = \tilde{\beta}_0 &+ (\beta_b \frac{\gamma_{bp}}{\gamma_b} + \beta_p) P_j' + (\beta_{b\bar{y}} \frac{\gamma_{bp}}{\gamma_b} + \beta_{p\bar{y}}) P_j' m(Y_{-ict-1}) + (2\beta_{b\bar{y}}\beta_b \frac{\gamma_y}{\gamma_b} + \beta_{\bar{y}}) m(Y_{-ict-1}) + \\
&+ \beta_{b\bar{y}}^2 \frac{\gamma_y}{\gamma_b} m(Y_{-ict-1})^2 + (\beta_y + 2\beta_{by}\beta_b \frac{\gamma_y}{\gamma_b}) Y_{it-1} + \beta_{by}^2 \frac{\gamma_y}{\gamma_b} Y_{it-1}^2 + (\beta_{by} \frac{\gamma_{bp}}{\gamma_b} + \beta_{py}) P_j' Y_{it-1} + \\
&+ 2\beta_{by}\beta_{b\bar{y}} \frac{\gamma_y}{\gamma_b} Y_{it-1} m(Y_{-ict-1}) + \epsilon_{it}, \\
= &\alpha_0 + \alpha_p P_j' + \alpha_{p\bar{y}} P_j' m(Y_{-ict-1}) + \alpha_{\bar{y}} m(Y_{-it-1}) + \alpha_{\bar{y}2} m(Y_{-ict-1})^2 + \alpha_y Y_{it-1} + \alpha_{y2} Y_{it-1}^2 \\
&+ \alpha_{py} P_j' Y_{it-1} + \alpha_{y\bar{y}} Y_{it-1} m(Y_{-ict-1}) + \epsilon_{it}.
\end{aligned} \tag{2}
$$

Note that the interaction between teaching practice and peer achievement is driven by two channels that are beyond the simple interaction included in the production function. First, student optimal behavior varies with the teacher's practice (i.e. she makes bad behavior more costly ($\beta_{bp} > 0$)). Second, the benefits of good behavior for achievement are higher if peer initial achievement is higher (i.e. the students can benefit from peers because they are behaving). In summary, it is important to highlight that the same reduced form expression can be obtained even when teaching practices are not included in the learning production function, so that $\beta_p = \beta_{py} = \beta_{p\bar{y}} = 0$ (i.e. an indirect effect will emerge through students endogenous responses).

While the model above 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 behavior 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.,

$$Y_{it} = \beta_0 + \beta_b b_{it} + \beta_{by} b_{it} Y_{it-1} + \beta_{b\bar{y}} b_{it} m(Y_{-ict-1}) + \beta_{b\bar{b}} b_{it} \bar{b}_{-it} + \beta_{\bar{b}} \bar{b}_{-it} + \beta_y Y_{it-1} + \beta_{\bar{y}} m(Y_{-ict-1}) + \beta_p P_j' + \beta_{py} P_j' Y_{it-1} + \beta_{p\bar{y}} P_j' m(Y_{-ict-1}) + \epsilon_{it},$$

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 $m(Y_{-ict-1}) = \bar{Y}_{-ict-1}$, 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.

The model so far takes as given the teaching practice, whereas in reality teachers could respond to the classroom composition by modifying their teaching practice. As we discuss below, we think we can identify most convincingly the effects of a fixed or persistent aspect of teaching practice and so do not focus on this channel.

4 Estimation

Our empirical strategy focuses on estimation of the reduced form model described in equation (2), as this if the focus of the literature. We take as a starting point that $m(Y_{-ict-1}) = \bar{Y}_{-ict-1}$ and

expand to consider other moments of the peer achievement distribution in the application, i.e.,

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

where we assume that observed achievement is a result of student's utility-maximizing behavior. The parameter, $\alpha_{p\bar{y}}$, which captures how the marginal benefits of teaching practice vary with the classroom composition constitutes our main object of interest. To simplify exposition, we ignore the role of student and teacher observables (e.g. race, gender, among others) though we show results that include these covariates.

As discussed above, a unique aspect of these data is that teachers are randomly assigned to classrooms. This means that teacher fixed and pre-determined characteristics are independent of ϵ_{it}. Because randomization held within randomization blocks at a school, we control for randomization block fixed effects to isolate this random variation. 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, which we need to address. Largely, this was because assignments were made based off of preliminary rosters at the end of the previous school year 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. This means that $\bar{Y}_{-ict-1}$ could be correlated with the student's unobserved type for instance, even though it is not correlated with unobserved teacher quality. Third, P_j may still be endogenous even with random assignment because of measurement error. We discuss each of these issues in turn.

Non-compliance We index the randomly assigned teacher as $r = r(i,t)$, so that the teaching practice of the randomly-assigned teacher is denote P_r. 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, instrumenting for teaching practice of the observed teacher using the teaching practice of the randomly assigned teacher. However, in our setting we need to instrument

for the level effect of the practice and its two interactions. This is easily done with the additional instruments $P_r\bar{Y}_{-ict-1}$ and P_rY_{-it-1}. However, in our main models when we include both practices and multiple moments of peer initial achievement, this introduces considerable noise. Therefore, we focus much of our discussion around the "intent-to-treat" estimates, which replace the observed teaching practice with the randomly-assigned teaching practice. The benefit of this latter strategy is that it has smaller standard errors, but the cost is that it understates the benefits of teaching practice, as we show in Section 5.

Because teachers were randomly assigned at the randomization block levels, we include randomization block fixed effects α_b, where $b = b(i,t)$ indexes randomization blocks. Thus, our equation 3 becomes

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

Endogeneity of classroom composition Endogeneity of classroom composition could occur in our data for two reasons. First, the principals were not required to assign classroom composition randomly. Second, non-compliance by students could lead the classroom composition to be endogenous even after addressing the non-compliance at the teacher-level.

The question we now address is whether we can identify $\alpha_{p\bar{y}}$ even though $\bar{Y}_{-ict-1}$ is potentially endogenous.[16] Maintaining the assumption that practice does not respond to classroom composition, we have that ϵ_{it} is independent of R_t. For simplicity, we ignore for the moment the conditioning on Y_{it-1} and randomization block fixed effects, though all arguments go through with this additional conditioning.[17] Assume further without loss of generality that $E(P_r) = E(\bar{Y}_{-it-1}) = 0$, so that teaching practice and classroom composition measures are mean 0. Demeaning these variables also aids in interpretation of the parameters in equation (4) as discussed further in Section 5.

The correlation between the interaction term and the residual can be written as $Cov(P_r\bar{Y}_{-it-1}, \tilde{\epsilon}_{it}) =$

[16] Bun and Harrison (2014) and Nizalova and Murtazashvili (2014) provide a detailed discussion of this type of setting, where an exogenous covariate is interacted with an endogenous variable, which we follow here.

[17] We also ignore the higher order peer terms though inclusion of them does not change our results.

$E(P_r E(\bar{Y}_{-it-1}\epsilon_{it}|P_r)$. Sufficient assumptions for identification of the interaction include

Assumption A1. $E(\bar{Y}_{-ict-1}\epsilon_{it}|P_r) = E(\bar{Y}_{-ict-1}\epsilon_{it})$, *and*

Assumption A2. $\bar{Y}_{-ict-1}$ *is independent of* P_r

Thus, for instance, if there is matching of students to peers which generates a correlation between peer initial achievement and the residual, it is independent of the randomly assigned teaching practice.

It then follows that $Cov(P_r\bar{Y}_{-ict-1}, \epsilon_{it}) = E(P_r)E(\bar{Y}_{-ict-1}\epsilon_{it}) = 0$, given that $E(P_r) = 0$. The first equality follows from random assignment of teachers to students and the second through a normalization of the independent variables, without loss of generality.[18] Bun and Harrison (2014) and Nizalova and Murtazashvili (2014) show that the independence between P_r and $\bar{Y}_{-ict-1}$ and that the covariance of $\bar{Y}_{-ict-1}$ and $\tilde{\epsilon}_{it}$ does not vary by P_r created by random assignment are sufficient to obtain unbiased estimates of $\alpha_{p\bar{y}}$ even if $\bar{Y}_{-ict-1}$ is endogenous. Nizalova and Murtazashvili (2014) point to different studies using randomized control trials that maintain this assumption when estimating heterogeneity in treatment effects. Bun and Harrison (2014) point out that a number of weaker versions of Assumption A2 are sufficient for identiifcation. In particular, it would be sufficient if $E(\bar{Y}_{-ict-1}P_r) = E(\bar{Y}_{-ict-1})E(P_r)$ and $E(\bar{Y}^2_{-ict-1}P_r) = E(\bar{Y}^2_{-ict-1})E(P_r)$.

The main way assumptions A1 and A2 could be violated is if students reshuffle after the initial teacher assignment. We do not believe this is a concern for 2 reasons. First, as we discussed in the balancing tests in Section 2, we do not see evidence that peer characteristics (measured after potential reshuffling of students) are correlated with teaching practice. Second, we test this condition by regressing each of the student's own characteristics times the peer characteristics on the randomly assigned teaching practice. These tests also provide strong support that at least in terms of observables this condition is not violated. Finally, we can test the implications for our estimation if there is some matching based on unobservables that we did not detect with our tests, by replacing the observed peer characteristics with the initially-assigned peer characteristics in our regressions. We will show that results are robust to this setting in Section 5.

[18] It is important to emphasize that in our ITT specifications if a student is ex-post re-allocated to a different classroom, we still impose that the teaching practice that he/she is exposed corresponds to the randomly assigned teacher of his original classroom.

Measurement error and endogeneity of teaching practice Suppose that true teaching practice (P_j) is measured with error. 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. Let the subscript k capture different observations of the teaching practice, i.e.,

$$P_{jkt} = \delta_k P_j + u_{jkt}. \tag{5}$$

Our preferred approach is to use $t-1$ measures to capture the teaching practice. This address two related concerns. First, video raters may have difficulty separating the teacher's practice from the students they are teaching. In fact, a cursory look at the protocol descriptions suggests that this may be an important challenge. Second, if teachers change their practice in response to classroom composition, then the practice of the randomly assigned teacher is no longer independent, leading our identification strategy to fail.

To clarify the potential effects of measurement error on our estimates, suppose we replace the true teaching practice with one of the measures of teaching practice in $t-1$ assuming that teaching practice is a scalar, i.e.,

$$Y_{it} = \alpha_0 + \frac{\alpha_p}{\delta_k} P^j_{rkt-1} + \frac{\alpha_{p\bar{y}}}{\delta_k} P^j_{rkt-1}\bar{Y}_{-ict-1} + \alpha_{\bar{y}}\bar{Y}_{-ict-1} + \alpha_{\bar{y}2}\bar{Y}^2_{-ict-1} + \alpha_y Y_{it-1} + \alpha_{y2}Y^2_{it-1} + \frac{\alpha_{py}}{\delta_k} P_{rkt-1}Y_{it-1} + \alpha_{y\bar{y}}Y_{it-1}\bar{Y}_{-ict-1} + \alpha_b + \nu_{it}$$

where $\nu_{it} = \tilde{\epsilon}_{it} - \frac{\alpha_p}{\delta_k} u_{rkt-1} - \frac{\alpha_{p\bar{y}}}{\delta_k} u_{rkt-1}\bar{Y}_{-ict-1} - \frac{\alpha_{py}}{\delta_k} u_{rkt-1}Y_{it-1}$. If measurement error is random, this should bias our estimates of α_p and $\alpha_{p\bar{y}}$ toward 0. If $Cov(u_{rkt-1}, P_{rkt-1}) > 0$, which is expected if there is assortative matching of better teachers with better students in $t-1$, then estimates of $\alpha_{p\bar{y}}$ would be further biased toward 0.

The primary results we focus on use simple averages across relevant observations of our measures of practice. However, we show results are robust to using principle component analysis to construct our measures (the primary approach we have seen applied in this literature) or factor model 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. 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 C.

To the extent that practice is time-varying, the focus on $t-1$ measures may understate the total effect of teaching practice though the potential effect on our interaction terms is ambiguous. For time-varying practice, we can extract instead the common component from the correlation between $t-1$ and time t practices, as this should not be related to classroom composition, given that teachers are randomly assigned to classrooms in time t. This would pick up a persistent aspect of teaching practice. Because we need to deal with non-compliance as well, this means instrumenting for the time t observed teacher's teaching practice with the time $t-1$ teaching practice of the randomly assigned teacher, which introduces more noise.

5 Results

We begin by showing results from simple specifications that ignore the potential complementarities of teaching practice and classroom composition, and then we build into more complex models. This relates most closely to the existing literature and will clarify why the simpler models that are often the focus of the literature can provide misleading evidence on the benefits of teachign practice. Furthermore, this will show how teaching practices emerge as important elements of the learning process once interactions with classroom composition are brought into the analysis.

5.1 Do Teaching Practices have a Direct Effect on Test Scores?

We begin by estimating a simpler version of equation (3), where we abstract from key interactions between teaching practices and peer characteristics. The aim is to study whether these practices play any visible role when we do not account for complementarities with classroom composition. Panels A and B of Table 3 display results from eight different specifications for behavior management and student-centered practices, respectively. In particular, odd columns present results from models without any relevant interaction, while even columns additionally incorporate interactions between student prior performance and the teaching practice of interest. Results in columns (1)

and (2) correspond to naive OLS specifications, where previous year teaching practice of the current teacher (P_{jt-1}) is the variable of interest.[19] Columns (3) and (4) report intent-to-treat (ITT) estimates, replacing P_{jt-1} with the teaching practice at $t-1$ 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}. Finally, columns (7) and (8) also report IV findings but in this case P_{rt} is instrumented with P_{rt-1} to capture persistent aspects of teaching practice across years in the case where teaching practice is time-varying.[20]

Given the breadth of the measures, it is perhaps suprising that none of the specifications (in both panels) show that the level of teaching practices play a statistically significant role in math performance.[21] However, these results are consistent with the findings in Garrett and Steinberg (2015), where principal components of FFT measures do not seem to have a direct impact on students' performance. In a similar vein, while interactions of student prior achievement with behavior management or student-centered practices are statistically significant in ITT and IV specifications, F-tests (reported at the bottom of each panel) show that the coefficients associated with these practices are in many specifications not jointly significant. At first glance, these findings suggest that our constructs of teaching practice may not measure something meaningful or at the very least do not matter for performance, and therefore they should be disregarded as relevant measures of the teacher's effectiveness in teaching math. Moreover, while peer effect parameters are likely biased upward (if anything) because of non-random sorting of students, we find that in all specifications average peer prior achievement has small coefficient estimates of the order of magnitude of 0.013 at the highest and large standard errors. Thus, these findings might also suggest that peer initial achievement does not substantially affect math performance. However, the following section shows that these conclusions are misleading when we build to account for complementarities between teaching practice and peers.

[19] Notice that we have access to measures of teaching practice for the same teacher at two different points in time (i.e. academic years 2009-2010 and 2010-2011). We do not report OLS results that include current teacher practices because results would be largely endogenous, though we report IV estimates that instrument teaching practices of the current randomly assigned teacher with its $t-1$ teaching practices.

[20] The first stage shows that P_{rt-1} is a statistically significant predictor of P_{rt} with a coefficient of 0.32 for behavior management and 0.17 for student-centered practices.

[21] These results also holds if instead of using averages of the sub-domains, we consider a principal component approach.

Table 3: Effects of Teaching Practice without Classroom Interactions

	Actual Teacher		Random Teacher		IV Actual with Rand. Teacher		IV Practice$_t$ With Practice$_{t-1}$ for Rand. Teacher	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A								
Behavior Management	0.012	0.012	0.015	0.015	0.017	0.016	0.046	0.046
	(0.021)	(0.020)	(0.019)	(0.018)	(0.021)	(0.020)	(0.066)	(0.064)
B.M.×Math$_{t-1}$		0.021*		0.022*		0.023*		0.041*
		(0.012)		(0.012)		(0.012)		(0.022)
Math$_{t-1}$	0.752***	0.752***	0.752***	0.752***	0.752***	0.752***	0.753***	0.753***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Avg Peer Math$_{t-1}$	0.012	0.013	0.011	0.012	0.012	0.013	0.0016	0.0043
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.030)	(0.029)
P-value (joint signif. of teaching practice)		0.210		0.168		0.151		0.150
F-Stat. (first stage)†					765.6	387.6	10.47	5.169
Panel B								
Student Centered	0.022	0.021	0.022	0.021	0.025	0.023	0.127	0.132
	(0.019)	(0.019)	(0.019)	(0.018)	(0.021)	(0.021)	(0.135)	(0.136)
S.C.× Math$_{t-1}$		0.017		0.023*		0.024**		0.068*
		(0.012)		(0.012)		(0.012)		(0.036)
Math$_{t-1}$	0.752***	0.752***	0.752***	0.751***	0.752***	0.751***	0.755***	0.753***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.018)
Avg Peer Math$_{t-1}$	0.011	0.010	0.010	0.008	0.011	0.009	0.001	0.003
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.028)	(0.026)
P-value (joint signif. of teaching practice)		0.142		0.048		0.038		0.154
F-Stat. (first stage)†					921.5	477.2	4.200	2.058

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 in columns (1)-(6) and sample size is 3322. Columns (7) and (8) use current practices and sample is 3319. 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 student characteristics listed in Table 1. † Reports the Kleibergen-Paap rk Wald statistic for a weak instrument test.

5.2 Teaching Practice and Classroom Composition

We expand the previous analysis by fully estimating equation (3), by including interactions of different moments of the classroom prior achievement distribution (i.e. average peer achievement, and interquantile range, IQR) with teaching practices. Panels A and B of Table 4 present results from six different specifications for each practice, respectively. Odd columns accommodate models where average peer prior achievement is interacted with one of the broad teaching domains (in addition to student prior achievement), while even columns additionally control for classroom interquartile range and its interaction with teaching practices. 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}. Finally, columns (5) and (6) provide further ITT robustness checks where we additionally replace the average and IQR of the actual classroom composition with those of the original composition (i.e. before any potential re-allocation of students could occur in response to the randomization of teachers into classrooms).[22]

Panel A shows that classrooms with higher average peer initial achievement benefit more from behavior 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 behavior management increase test scores around 5% to 3.5% of a standard deviation when peer average prior year performance is one standard deviation above the mean.[23] In contrast, the even columns show that the effectiveness of behavior management practices does not vary significantly with the IQR in classroom initial achievement. Furthermore, consistent with the results in Table 3, the level effects of behavior management practices are still not statistically significantly different from 0 and point estimates are small. Moreover, the interactions between behavior management and student's prior achievement become statistically insignificant in most specifications, suggesting that failure to account for complementarities with classroom composition may lead to stronger conclusions

[22] Recall that our identification strategy only requires the random allocation of teachers into classrooms to estimate the interaction between classroom composition and teaching practice, so is robust to the endogeneity of classroom composition. It is important to clarify that if a student is ex-post re-allocated to a different classroom, we impose that the teaching practice that he/she was exposed corresponds to the randomly assigned teacher.

[23] For example, our findings indicate that teaching practices that prevent misconduct seem to reinforce peer effects in the classroom.

about student-level heterogeneity in the effects of teaching practice. A further notable change is that behavior management emerges as a statistically significant predictor of test performance when interacted with average peer prior achievement. In addition, F-tests show that the variables associated with this teaching domain are jointly significant at the 99% confidence level in most specifications. The fact that the interaction between student prior achievement and behavior practice is not significant suggests that the channel in which teaching practices are operating is through amplifying the role of peer effects, as discussed in section 4, and that failure to account for this interaction would lead us to misrepresent the relevance of this teaching domain for math performance.

Panel B shows results for student-centered practices. Generally, we find that classes with higher average initial achievement also benefit more for student-centered practices. However, the benefits of student-centered practices are smaller in classroom with higher IQR in initial achievement. Like in the case of behavior management, the level effect of student centered practices are not statistically significantly different from 0 and neither are the interactions with initial achievement, after controlling for interactions with classroom composition. Furthermore, joint tests also confirm that student-centered practices are significant predictors of achievement, with a p-value of 0.003 in our preferred specification. Finally, there may be some concerns that IQR may be correlated with class size, and therefore we may be picking class size effects. However, specifications that control for class size and its interactions do not change the results. In summary, the findings in Table 4 provide two main messages. First, teaching practices seem to show large complementarities with classroom characteristics. Second, the contrasting evidence between behavior management and student-centered practices also points to the importance of considering these measures separately, i.e., a single measure of teaching quality does not allow to understand in which contexts different practices become more or less effective.

A valid concern with our findings is to what extent our results (e.g. lack of significance in the level of the teaching practice measures) are affected by problems of measurement error in our key teaching practice variables. In order to address this point, we implemented a measurement error correction strategy that follows Hausman et al. (1991). This approach is more convenient than

Table 4: Teaching Practice and Classroom Composition

	Random Teacher		IV Actual with Rand. Teacher		IV Practice$_t$ With Practice$_{t-1}$ for Rand. Teacher	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A						
Behavior Management	0.017	0.020	0.018	0.021	0.068	0.068
	(0.016)	(0.017)	(0.018)	(0.019)	(0.062)	(0.062)
B.M. $\times$ Math$_{t-1}$	0.014	0.013	0.015	0.013	0.029	0.025
	(0.012)	(0.011)	(0.012)	(0.012)	(0.022)	(0.021)
B.M. $\times$ Avg Peer Math$_{t-1}$	0.049***	0.048**	0.054***	0.052**	0.140**	0.125**
	(0.018)	(0.018)	(0.019)	(0.021)	(0.057)	(0.064)
B.M. $\times$ IQR Peer Math$_{t-1}$		-0.015		-0.015		-0.032
		(0.016)		(0.019)		(0.038)
P-value (joint signif. of teaching practice)	0.010	0.008	0.004	0.002	0.052	0.029
First Stage F-Stat.†			210.9	87.94	3.371	2.981
Panel B						
Student Centered	0.019	0.012	0.020	0.011	0.222	0.134
	(0.019)	(0.019)	(0.022)	(0.021)	(0.250)	(0.225)
S.C. $\times$ Math$_{t-1}$	0.016	0.0130	0.017	0.013	0.049	0.038
	(0.011)	(0.011)	(0.012)	(0.011)	(0.034)	(0.033)
S.C. $\times$ Avg Peer Math$_{t-1}$	0.032**	0.030**	0.036**	0.034**	0.195	0.191
	(0.015)	(0.0150)	(0.017)	(0.017)	(0.170)	(0.165)
S.C. $\times$ IQR PeerMath$_{t-1}$		-0.035**		-0.039**		-0.096
		(0.015)		(0.016)		(0.071)
P-value (joint signif. of teaching practice)	0.015	0.010	0.009	0.003	0.500	0.365
First Stage F-Statistic†			151.4	75.22	0.521	0.418

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 3322. Lagged teaching practices are used with the exception of columns (5) and (6), where sample size is 3319. 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 student characteristics listed in Table 1. 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.

the usual IV strategy that accounts for error in variables, because in our model the variables of interest enter non-linearly into the model. In appendix C, we provide a description of how we adapt the Hausman et al. (1991) method to our context, and 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 alternative approaches.

5.3 Teaching Practice vs. Teacher "Quality"

While previous specifications provide important insights, teachers who have better behavior management practices may also engage in more student-centered practices; therefore not including both domains in the same specification may bias our estimates. In addition, a remaining concern is whether our findings on complementarities between teaching practices and peer composition are robust to the inclusion of measures of teacher "quality." It is worth pointing out that 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 behavior management and student centered practices are in fact measures of quality. Furthermore, the fact that behavior-management and student-centered practices interact differently with classroom composition already suggests that a single unidimensional quality may not be correct. That said, it is still informative to see whether controlling for more traditional measures of teacher quality changes our results.

In order to address these key points, Table 5, 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 behavior management and student-centered practices and their interactions with peer composition. These results show that interactions of behavior management with the average peer initial achievement are robust, but seem to explain the interaction of 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 student-centered practices with the

IQR in peer initial achievement remain robust.

Columns (3) to (5) of Table 5 report results from ITT specifications similar to column (1) where we additionally include different proxies for overall teaching "quality" and its interactions with classroom characteristics.[24] First, as a measure of aptitude we include teacher performance in the Content Knowledge for Teaching (CKT) assessment, which has been designed to measure the extent to which a prospective elementary teacher has the content knowledge needed in the areas of reading and language arts, mathematics, science, and social studies.[25] Second, we include the teacher's average score on student assessments from the TRIPOD survey. This instrument assesses the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth.[26] Finally, we included school principal evaluations on teachers performance which are reported in the MET database.[27] These results show that across all specifications our key interactions between teaching practices and moments of the classroom prior achievement distribution remain significant where the size of these coefficients is fairly constant, and similar to our previous specifications. Furthermore, we see that these alternative measures of "quality" do not interact with peer average initial achievement and IQR in the same way as our two practices. This is true despite CKT and principal surveys being statistically significant predictors of math achievement. 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 better students.

[24] Notice that we cannot control for the usual measures of teacher value-added (i.e. adjusted random effects) because these models inherently neglect the presence of classroom-teacher interactions.

[25] See MET project (2010a) for a detailed description of this assessment.

[26] Tripod is a research-based 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.

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

Table 5: Teaching Practices and Alternative Teacher "Quality" Controls

	Random Teacher	IV Actual with Random Teacher	Random Teacher Alt. Teacher Control:		
			CKT	7C	PSVY
	(1)	(2)	(3)	(4)	(5)
Behavior Management	0.007	0.007	-0.004	0.011	0.005
	(0.023)	(0.026)	(0.022)	(0.023)	(0.022)
B.M. × Math_{t-1}	0.008	0.007	0.009	0.007	0.004
	(0.017)	(0.018)	(0.017)	(0.017)	(0.017)
B.M. × Avg Peer Math_{t-1}	0.045*	0.049*	0.055**	0.048*	0.043*
	(0.025)	(0.028)	(0.025)	(0.025)	(0.024)
B.M. × IQR Peer Math_{t-1}	0.018	0.023	0.013	0.021	0.014
	(0.022)	(0.025)	(0.022)	(0.022)	(0.022)
Student Centered	0.010	0.01	0.016	0.011	0.002
	(0.026)	(0.029)	(0.024)	(0.026)	(0.026)
S.C. × Math_{t-1}	0.008	0.008	0.003	0.015	0.008
	(0.017)	(0.018)	(0.017)	(0.017)	(0.016)
S.C. × Avg Peer Math_{t-1}	0.003	0.005	-0.006	-0.003	0.004
	(0.021)	(0.024)	(0.020)	(0.021)	(0.022)
S.C. × IQR Peer Math_{t-1}	-0.044**	-0.050**	-0.046**	-0.050**	−0.037
	(0.021)	(0.023)	(0.019)	(0.024)	(0.023)
Alt. Teacher Control			−0.014	−0.002	0.056***
			(0.016)	(0.019)	(0.016)
T.C.× Math_{t-1}			0.043***	-0.0230*	0.033***
			(0.013)	(0.012)	(0.011)
T.C.× Avg Peer Math_{t-1}			-0.020	0.02	-0.022
			(0.020)	(0.021)	(0.016)
T.C. × IQR Peer Math_{t-1}			-0.007	0.013	-0.013
			(0.020)	(0.023)	(0.016)
P-value joint signif of BM& SC	0.033	0.008	0.034	0.015	0.143
P-value joint signif. T.C.			0.030	0.406	0.002
First Stage F-Statistic†		31.84			

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 3322. 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 1. 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.

6 Conclusion

Exploiting a unique data set that provides detailed information on teaching practices from trained raters based on an increasingly popular teacher evaluation protocol, FFT, and where teachers are randomly assigned to classrooms, this paper makes three main contributions to the literature on teacher effectiveness. First, we show that the benefits of teaching practices vary significantly with classroom composition. This finding suggests that teachers can in fact shape peer effects and classroom composition can shape teacher effects. We find that behavior management practices are more effective in classes with higher average initial peer achievement, whereas student-centered practices are more effective in classrooms with lower variation in initial peer achievement.

Second, failure to capture the interactions of the teaching practice with the classroom composition may lead researchers to understate or even conclude that a teaching practice is not relevant for achievement when in fact it matters. In particular, we find that level effects of the teaching practices measured by FFT on math performance are not statistically significantly different from 0.

Third, failure to capture the interactions of the teaching practice with the classroom composition also may lead researchers to misstate the importance of heterogeneity in the effects of the teaching practice by the individual student's initial achievement. Both behavior management and student-centered practices seem to matter more for higher-achieving students before including interactions with classroom composition.

We illustrate through a simple model the potentially pervasive nature of the interactions of practice with classroom composition. For instance, the interaction of behavior management practices with average peer initial achievement arises in a model where the achievement benefits of good behavior (which is positively affected by behavior management practices) are higher if a student's peers are also behaving or if her classmates are higher-achieving. That student-centered practices are more effective in lower variance classrooms seems more straightforwardly a property of the achievement production function, though it could also be motivated through preferences if student engagement or effort depends on their initial achievement relative to their peers.

We show that our results are robust to a number of specification checks, including controlling for

observable unidimensional aspects of teacher quality, including aptitude from the Content Knowledge assessment, principal surveys of teacher quality and student surveys do not explain these patterns. Furthermore, similar effects are not found with these unidimensional measures of quality. These results suggest that the common focus in the teacher effectiveness literature, including the value-added literature, to treat teaching quality as a unidimensional, separable construct may be misguided.

Finally, we consider implications of these findings on classroom composition and teaching practice in two school improvement paradigms: (1) teacher evaluation and accountability, and (2) 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. When coupled with other indicators of teacher effectiveness, rigorous observations might help reduce the negative instructional adaptations that sometimes occur in test-based accountability systems (Jennings and Corcoran, 2012), and reduce the risk of erroneously labeling any given teacher as ineffective or effective. Yet, our findings suggest that, depending on teachers' assignments or the overall school 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.

Our main findings also have ramifications for advancing teacher training and professional development. In particular, they reinforce the importance of explicit attention to challenges stemming from classroom-achievement heterogeneity (Cohen and Lotan, 1997; Seaton et al., 2010). Another ramification is that behavioral management practices appear to be especially important in cultivating achievement growth in high-achieving elementary school classrooms. On the one hand, this is intuitive given that students cannot benefit from having higher-achieving peers if the teacher does not maintain a good classroom environment by managing behaviors. On the other hand, this finding is novel or even counter-intuitive given models of student motivation and engagement that stress the iterative relationship between achievement and engagement; i.e. problems of engagement and disruptive behaviors are often traced to academic struggles (Finn and Zimmer, 2012; Voelkl,

2012), and are thought to be exacerbated in low-achieving peer-contexts (Kelly, 2009). Yet, high-achieving students can also experience low levels of engagement, and in particular, low levels of concentration caused by a mismatch between challenge and skills (Shernoff et al., 2003). Given this interpretation, we suggest further research to better isolate the effects of teachers' emphasis on behavioral management from the accomplishment of a positive behavioral climate, which may stem from other underlying instructional conditions including levels of challenge.

Finally, returning to the need to differentiate teachers' practice in specific instructional domains in teacher evaluation, an ancillary impression from our analysis of the MET data arose which suggest limits to present capacity; scores on protocol subdomains do not appear to be as orthogonal in practice as they are in principle, or are intended to be. 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.[28] But at least in this study, domain scores are much more similar than they are disparate. Again, further research is needed on observing and measuring teachers' practice in order to more fully separate the specific aspects of teaching practice which are theorized to matter for achievement. Our results indicate that this needs to be done with attention to the important moderating effect of classroom composition.

[28]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).

References

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

Araujo, Maria Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady, "A Helping Hand? Teacher Quality and Learning Outcomes in Kindergarten," 2014. working paper.

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," School of Economics Working Paper Series 2014-3, LeBow College of Business, Drexel University January 2014.

Burke, Mary A. and Tim R. Sass, "Classroom Peer Effects and Student Achievement," Working Papers, Department of Economics, Florida State University February 2006.

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.

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.

Finn, Jeremy D. and Kayla S. Zimmer, "Student Engagement: What Is It? Why Does It Matter?," in Sandra L. Christenson, Amy L. Reschly, and Cathy Wylie, eds., *Handbook of Research on Student Engagement*, Boston, MA: Springer US, 2012, pp. 97–131.

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

Fruehwirth, Jane Cooley, "Identifying peer achievement spillovers: Implications for desegregation and the achievement gap," *Quantitative Economics*, 2013, *4* (1), 85–124.

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.

Gibbons, Steve and Shqiponja Telhaj, "Peer Effects and Pupil Attainment: Evidence from Secondary School Transition," CEE Discussion Papers 0063, Centre for the Economics of Education, LSE May 2006.

Hanushek, Eric A. and Steven Rivkin, "Harming the best: How schools affect the black-white achievement gap," *Journal of Policy Analysis and Management*, 2009, *28* (3), 366–393.

_ , John F. Kain, and Steven G. Rivkin, "New Evidence about Brown v. Board of Education: The Complex Effects of School Racial Composition on Achievement," *Journal of Labor Economics*, 2009, *27* (3), 349–383.

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.

Hoxby, Caroline M. and Gretchen Weingarth, "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects," 2005. Working Paper.

Jennings, Jennifer L. and Sean P. Corcoran, "Beyond high-stakes tests: Teacher effects on other educational outcomes," in Sean Kelly, ed., *Assessing teacher quality: Understanding teacher effects on instruction and achievement*, New York: Teachers College Press, 2012, pp. 77–96.

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.

Kelly, Sean, "Social identity theories and educational engagement," *British Journal of Sociology of Education*, 2009, *30* (449–462).

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

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.

MET project, "Content knowledge for teaching and the MET project," Bill and Melinda Gates foundation September 2010.

_ , "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.

Voelkl, Kristin E., "School Identification," in Sandra L. Christenson, Amy L. Reschly, and Cathy Wylie, eds., *Handbook of Research on Student Engagement*, Boston, MA: Springer US, 2012, pp. 193–218.

A Randomization and Sample Selection

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.[29] Kane et al. (2013)

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 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, 79% of classroom peers were included in randomization, and we observe prior test scores for 91% of classroom peers.

To motivate our sample selection, we first review what is needed for estimation. We rely on the random assignment of classes to teachers, and primarily use teaching practice from the prior year to avoid endogeneity issues. Additionally, we need current and prior test scores to measure the effect of teaching practice on a student's outcomes. We start with the entire sample of elementary students in the randomization year (2010-11), in either a math or joint math and ELA classroom and restrict the sample to students who were a part of random assignment. At this point we have 5,730 students, summarized in appendix Table (9). Next, we restrict the sample to randomization blocks in which half or more of the students complied with random assignment. We lose 31% of the sample here, but still have 3,927 student observations. One reason for this is that compliance rates and teacher attrition varied by school, but since our analysis relies on within randomization block variation in teaching practice and classroom composition, this does not affect the internal validity of our results. Next we restrict the sample to students we can use in our baseline specifications. This just requires that students have current and prior test scores, and that their actual and randomly assigned teacher has non-missing measures of teaching practice in the prior year (2009-10). We lose just 149 observations or under 4% of the remaining sample. Next we make a class size restriction,

[29]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.

requiring that all classes have a minimum of 7 students in random assignment with current and prior test scores. We do this to avoid the possibility of results being driven by unusually small classes. We only lose 38 students or about 1% of the remaining sample from this class size restriction directly. Our estimation strategy requires a minimum of two teachers per randomization block. We lose about 10% of the remaining sample, and are left with 3,372 students who show up in our analysis. There are 22 students showing up twice within schools but between sections. To avoid counting these students twice, we drop all of these duplicate student observations or 44 student observations. Finally, we check the randomization block compliance rates again, the class size restriction and that there are at least two teachers per randomization block. The final restricted regression sample has 3,322 student observations. These student observations span 5 districts, 45 schools, 70 randomization blocks, 183 teachers, 183 classrooms, with 87% of student observations coming from joint math/ela courses. Tables (1) and (9) present summary statistics of our sample before and after we make our main sample restrictions.

B Appendix Tables

Table 6: 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	0.010	0.018	0.022	0.019	0.018	0.012
	(0.017)	(0.020)	(0.025)	(0.016)	(0.020)	(0.025)
T.P. $\times$ Math$_{t-1}$	0.012	0.008	0.009	0.000	0.017	0.016
	(0.012)	(0.014)	(0.017)	(0.011)	(0.013)	(0.019)
T.P $\times$ Avg Peer Math$_{t-1}$	0.034*	0.048**	0.061**	0.016	0.039**	0.015
	(0.018)	(0.022)	(0.027)	(0.016)	(0.017)	(0.032)
T.P. $\times$ IQR Peer Math$_{t-1}$	-0.029*	-0.011	-0.021	-0.028*	-0.036**	-0.056**
	(0.017)	(0.017)	(0.025)	(0.015)	(0.016)	(0.024)
P-value joint signif. T.P.	0.025	0.141		0.179	0.005	
First Stage F-Stat.†		32.29			31.36	
Hansen J P-value††		0.405			0.539	
p2 load			1.062			0.836
p3 load			0.857			0.860

Notes: *** denotes significance at the 1%, ** at the 5% and * at the 10% levels. Sample size is 3322. The ITT columns uses randomly assigned MCP or ESL scores as "Practice." The IV columns use all other practices that load on behavior management to instrument for MCP, and likewise for ESL with 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 student centered practices as the third measurement since we are overidentified. The specification is identical to that in Table (4) 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 C. Standard errors are clustered at the randomization block level, and with the adapted Hausman estimator we bootstrap standard errors with 200 repetitions.

Table 7: Description of Framework for Teaching (FFT)

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

Table 8: 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 (7) for a detailed description of each FFT variable.

Table 9: Summary Statistics: Pre-Restricted Sample

	Mean	SD	Min	Max
Grade Level	4.515	0.50	4.00	5.00
Joint Math and ELA Class	0.850	0.36	0.00	1.00
Age	9.458	0.96	7.52	13.20
Male	0.490	0.50	0.00	1.00
Gifted	0.078	0.27	0.00	1.00
Special Education	0.090	0.29	0.00	1.00
English Language Learner	0.148	0.36	0.00	1.00
Free and Reduced Price Lunch Eligible	0.558	0.50	0.00	1.00
White	0.281	0.45	0.00	1.00
Black	0.345	0.48	0.00	1.00
Hispanic	0.273	0.45	0.00	1.00
Asian	0.071	0.26	0.00	1.00
American Indian	0.005	0.07	0.00	1.00
Race Other	0.022	0.15	0.00	1.00
Math Score (Year 09-10)	0.107	0.93	-3.14	2.84
Math Score (Year 10-11)	0.143	0.93	-3.26	3.02
Unique Districts	6	-	-	-
Unique Classes	362	-	-	-
Unique Schools	102	-	-	-
Unique Randomization Blocks	156	-	-	-
Unique Teachers	362	-	-	-
Percentage of Class w/ 09-10 Math Scores	0.908	0.07	0.63	1.00
Percentage of Class in Random Assignment	0.755	0.19	0.03	1.00
Teachers per Randomization Block	3.040	1.56	1.00	13.00
Randomization block compliance rate	0.662	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 10: Balance Tests

	Behav. Manag. Rand. Teach. (Year 2009-2010)	Student Centered Rand. Teach (Year 2009-2010)	Peer Math	IQR Math	Peer Math Rand	IQR Math Rand
Peer Math	0.021	0.086				
	(0.080)	(0.12)				
IQR Math	0.058	0.042				
	(0.091)	(0.090)				
Peer Math Rand	-0.021	0.073				
	(0.100)	(0.124)				
IQR Math Rand	-0.027	0.017				
	(0.078)	(0.075)				
Math_{t-1}	-0.0064	0.016	0.077*	-0.023	0.048	0.002
	(0.018)	(0.022)	(0.044)	(0.031)	(0.041)	(0.022)
ELL	-0.023	-0.017	-0.154	0.050	-0.149	-0.007
	(0.048)	(0.051)	(0.097)	(0.098)	(0.103)	(0.075)
Gifted	-0.016	-0.036	0.387**	0.131	0.212	0.218**
	(0.062)	(0.118)	(0.190)	(0.090)	(0.146)	(0.106)
Special Educ.	0.072	0.017	-0.123**	0.031	-0.065	0.007
	(0.052)	(0.065)	(0.060)	(0.065)	(0.056)	(0.052)
Male	0.004	0.001	-0.014	0.003	-0.023	-0.024
	(0.012)	(0.015)	(0.015)	(0.015)	(0.018)	(0.017)
White	0.014	-0.029	0.033	0.010	-0.031	-0.003
	(0.027)	(0.030)	(0.036)	(0.032)	(0.020)	(0.030)
Black	-0.010	-0.012	0.001	0.005	0.044*	0.017
	(0.025)	(0.029)	(0.040)	(0.046)	(0.023)	(0.043)
Hispanic	-0.044*	-0.036	-0.028	-0.010	-0.014	-0.009
	(0.025)	(0.030)	(0.025)	(0.040)	(0.024)	(0.040)
Asian	0.094*	0.141**	0.063*	-0.026	0.047	0.006
	(0.053)	(0.054)	(0.037)	(0.064)	(0.041)	(0.062)
American Indian	0.079	0.058	-0.232	0.209	-0.151	0.075
	(0.129)	(0.107)	(0.155)	(0.155)	(0.158)	(0.112)
Race Other	0.053	0.051	-0.078*	-0.028	-0.073**	-0.038
	(0.059)	(0.077)	(0.044)	(0.056)	(0.034)	(0.050)
R-squared	0.573	0.489	0.487	0.711	0.504	0.529
Obs.	3322	3322	3322	3322	3322	3322

Notes: We regressed each dependent variable separately on each independent variable with randomization block fixed-effects and stacked the parameters from these regressions. Behav. Manag. Rand. and Stud. Cent. Rand. are the averages of FFT components that load on factors one and two, respectively. "Rand." refers to a student's randomly assigned teacher in Year 10-11, but construct the variable using scores from Year 09-10.

C 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). Let $\tilde{Y}$ denotes the Y is demeaned at the randomization block level and similarly for other variables, then

$$\tilde{Y}_{it} = \alpha_p \tilde{P}_r + \alpha_{p\bar{y}} \widetilde{P_r \bar{Y}}_{-ict-1} + \alpha_{\bar{y}} \tilde{\bar{Y}}_{-ict-1} + \alpha_y \tilde{Y}_{it-1} + \alpha_{py} \widetilde{P_r Y}_{it-1} + \tilde{\epsilon}_{it}. \tag{6}$$

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 which was made for nonlinearities captured by polynomials in the variable of interest to our setting, where nonlinearities arise from interactions.

The parameters of equation (6) are identified from

$$E(\tilde{Y}_{it}) = \alpha_p E(\tilde{P}_r) + \alpha_{p\bar{y}} E(\widetilde{P_r \bar{Y}}_{-ict-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}) + \alpha_y E(\tilde{Y}_{it-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}) \tag{7}$$

$$\begin{aligned}
E(\tilde{Y}_{it}\tilde{P}_r) &= \alpha_p E(\tilde{P}_r\tilde{P}_r) + \alpha_{p\bar{y}} E(\widetilde{P_r\bar{Y}}_{-ict-1}\tilde{P}_r) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}\tilde{P}_r) + \alpha_y E(\tilde{Y}_{it-1}\tilde{P}_r) \\
&\quad + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{P}_r) \\
E(\tilde{Y}_{it}\tilde{\bar{Y}}_{-ict-1}) &= \alpha_p E(\tilde{P}_r\tilde{\bar{Y}}_{-ict-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r\bar{Y}}_{-ict-1}\tilde{\bar{Y}}_{-ict-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}\tilde{\bar{Y}}_{-ict-1}) + \alpha_y E(\tilde{Y}_{it-1}\tilde{\bar{Y}}_{-ict-1}) \\
&\quad + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{\bar{Y}}_{-ict-1}) \\
E(\tilde{Y}_{it}\widetilde{P_r\bar{Y}}_{-ict-1}) &= \alpha_p E(\tilde{P}_r\widetilde{P_r\bar{Y}}_{-ict-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r\bar{Y}}_{-ict-1}\widetilde{P_r\bar{Y}}_{-ict-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}\widetilde{P_r\bar{Y}}_{-ict-1}) \\
&\quad + \alpha_y E(\tilde{Y}_{it-1}\widetilde{P_r\bar{Y}}_{-ict-1}) + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\widetilde{P_r\bar{Y}}_{-ict-1}) \\
E(\tilde{Y}_{it}\tilde{Y}_{it-1}) &= \alpha_p E(\tilde{P}_r\tilde{Y}_{it-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r\bar{Y}}_{-ict-1}\tilde{Y}_{it-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}\tilde{Y}_{it-1}) + \alpha_y E(\tilde{Y}_{it-1}\tilde{Y}_{it-1}) \\
&\quad + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\tilde{Y}_{it-1}) \\
E(\tilde{Y}_{it}\widetilde{P_r Y}_{it-1}) &= \alpha_p E(\tilde{P}_r\widetilde{P_r Y}_{it-1}) + \alpha_{p\bar{y}} E(\widetilde{P_r\bar{Y}}_{-ict-1}\widetilde{P_r Y}_{it-1}) + \alpha_{\bar{y}} E(\tilde{\bar{Y}}_{-ict-1}\widetilde{P_r Y}_{it-1}) + \alpha_y E(\tilde{Y}_{it-1}\widetilde{P_r Y}_{it-1}) \\
&\quad + \alpha_{py} E(\widetilde{P_r Y}_{it-1}\widetilde{P_r Y}_{it-1})
\end{aligned}$$

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 subcomponent, 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

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

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 $\frac{\tilde{P}_{r2t}}{\delta_2}$, $\frac{\widetilde{P_{r2t}\bar{Y}}_{-ict-1}}{\delta_2}$ 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've 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 the robustness to using different measures as our anchor.

Appendix Table XX 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 behavior 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. behavior

management or student centered) lead to similar results that when we correct for measurement error by following other methods.