School-to-School Differences in Instructional Practice: New Descriptive Evidence on Opportunity to Learn

< Accepted for Publication (December 2020) in *Teachers College Record* >

by

Sean Kelly
(University of Pittsburgh)

Zachary Mozenter
(Washington, DC)

Esteban M. Aucejo
(Arizona State University)

Jane Cooley Fruehwirth
(University of North Carolina)

Please direct correspondence to Sean Kelly, 5527 Posvar Hall, 230 South Bouquet Street, Pittsburgh, PA 15260. Email: spkelly@pitt.edu. This research was supported by a grant from the Institute for Education Sciences (R305A170269). Any opinions, findings, and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of IES.

Structured Abstract

Background/Context: There is continuing debate among social scientists and educators about the role of school-to-school differences in generating educational inequality. Are some students high achieving because they attend School A, while others struggle because they attend School B, as critical discourse on schools argues? Alternatively, is educational inequality driven largely by social forces outside of the school, in the home and neighborhood environment, or by educational processes that are largely common across schools as much social science research argues? Analyses of school achievement, and in particular test score gains from year-to-year, suggest very small between-school differences. Yet, analyses of test score data alone may fail to reveal important school-to-school differences that affect the quality of the classroom experience and a variety of educational outcomes.

Purpose/Objective: We provide evidence on the following research questions. What is the magnitude of school-to-school variation in instructional practice, as captured by multiple measures? Are some domains of instruction (e.g. behavioral management) more variable between-schools than others? To what extent are school-to-school differences in instruction associated with compositional characteristics of students and teachers?

Research Design: This study relies on the Measures of Effective Teaching Study data, which offer an unprecedented set of observations of teachers' instruction scored on state of the art observational protocols. To examine the extent of school-to-school variation in instructional practice in elementary and middle schools, we conducted a decomposition of variance analysis using summary scores on multiple measures. We further examine behavioral climate as revealed during instruction separately from overall instructional practice. Next, we examine

differences in instruction associated with compositional characteristics of students using multilevel models. Finally, we utilize an innovative two-stage statistical adjustment strategy to more narrowly identify the possible association between composition and teaching practice due to school-to-school teacher sorting.

Findings/Results: The basic descriptive results from this study suggest a middle-view of school-to-school differences in instruction. We find that substantial school-level variation in instruction exists, with 30% or more of the total variance in instruction lying between schools in these data. Behavioral climate during instruction appears to be particularly salient, and especially in elementary schools. Much of the between school variance we identify, in some cases 40% or more, is readily explained by simple measures of socio-demographic composition, including in particular the racial make-up of schools in the MET districts. Finally, some evidence from a statistical adjustment method suggests that teacher sorting, rather than measurement bias and teacher adaptation is principally responsible for school-to-school differences in instruction.

Conclusions/Recommendations: More than an academic debate, basic differences between schools in the quality of the learning environment, along with parental understandings and beliefs about school effects, are potentially important drivers of school and neighborhood sorting and segregation, and even public investment in schooling. Additionally, this question carries continued policy relevance as states adopt and revise teacher and school accountability frameworks that implicitly attribute school-to-school differences to organizational functioning, and seek to carry out instructional improvement efforts in targeted schools. The basic descriptive results from this study suggest school-level differences are not as great as suggested by critical theory and the public discourse, but nor are they as inconsequential as one might

infer from some social science research or the literature on value-added differences between schools.

Executive Summary

Social scientists have long debated the role of school-to-school differences in generating educational inequality. Critical paradigms in the educational sciences hypothesize that many students are high achieving because they attend School A, while others struggle because they attend School B. Alternatively, is educational inequality driven largely by social forces outside of the school, in the home and neighborhood environment, or by educational processes that are largely common across schools?

In this study we offer a basic conceptual model of school effects, heavily influenced by Gamoran et al.'s updated nested layers model, positing that instructional variation across schools is created by three inter-related social processes: student and teacher sorting, school organizational functioning, and collective teacher adaptation. Importantly, collective teacher adaptation might entail forms of maladaptation to student background and classroom composition, including both over-adjustment of curriculum and negative responses to challenging behavior or interpersonal dynamics. Thus, it would not be surprising to find substantial school-to-school differences in instructional quality even as state and local policies seek to ameliorate such inequality.

Empirically, research on teacher and student sorting processes suggests that indeed, school segregation in its various forms creates tangible and significant differences in opportunity to learn between schools. In contrast, both basic research on school effects and applied studies of value-added in state administrative data suggest surprisingly weak variation in school

effectiveness. Based on test-score data alone, it seems that a large proportion of schools would best be described as "about average."

Yet, analyses of test score data alone may fail to reveal important school-to-school differences that affect the quality of the classroom experience, students' non-cognitive development, and long-range educational outcomes. Might more direct measures of the school experience reveal greater differences in opportunity to learn than research to date? Research on teacher effectiveness suggests that direct measures of instruction, student experience, and teacher knowledge and preparation yield valuable insight into the quality of students' educational experience.

In this study we turn to the Measures of Effective Teaching study data, which provides rich observational video data on instruction, scored using multiple well-developed protocols by highly-trained raters, as well as student reports of instructional practice, and measures of teacher knowledge. Our overall goal is to provide a descriptive portrait of school-to-school differences in instructional practice capitalizing on the great breadth of dependent measures of what happens in classrooms available in MET. We examine several specific research questions. What is the magnitude of school-to-school variation in instructional practice, as captured by multiple measures? Are some domains of instruction (e.g. behavioral management) more variable between-schools than others? To what extent are school-to-school differences in instruction associated with compositional characteristics of students and teachers?

The basic descriptive results from this study suggest a middle-view of school-to-school differences in instruction. We find that substantial school level variation in instruction exists, with 30% or more of the total variance in instruction lying between schools in these data. Behavioral climate during instruction appears to be particularly salient, and especially in

elementary schools. Much of the between school variance we identify, in some cases 40% or more, is readily explained by simple measures of socio-demographic composition, including in particular the racial make-up of schools in the MET districts. Finally, some evidence from a statistical adjustment method suggests that teacher sorting, rather than measurement bias and teacher adaptation is principally responsible for school-to-school differences in instruction.

The rigorous and extensive classroom observations in the MET study give us a view behind classroom doors seldom seen by most parents, or indeed even most educators. Educational research in the critical theory paradigm argues that pronounced differences in opportunity to learn between schools is an important driver of educational inequality. Similarly, much of the public discourse on school-to-school differences, the importance of choosing the "right" school, etc., indicate that beliefs about strong school effects are widespread. The basic descriptive results from this study suggest that school-level differences are not as great as suggested by critical theory and the public discourse, but nor are they as inconsequential as one might infer from school effects research, or in particular, the literature on value-added differences between schools. Of particular policy importance, we find that instructional quality, especially in elementary school, is associated with compositional features of schools, supporting critical theory arguments that student and teacher sorting practices create important differences in opportunity to learn.

Introduction

At least since the Coleman report (1966) on equality of educational opportunity, social scientists have debated the role of school-to-school differences in generating educational inequality. Are some students high achieving because they attend School A, while others

struggle because they attend School B (Bryk, Lee, & Holland, 1993; Carter, 2016; Demereth, 2007; Harris & Larsen, 2016)? Alternatively, is educational inequality driven largely by social forces outside of the school, in the home and neighborhood environment, or by educational processes that are largely common across schools (Connell et al., 1982; Downey & Condron, 2016)? More than an academic debate, basic differences between schools in the quality of the learning environment along with parental understandings and beliefs about school effects are potentially important drivers of school and neighborhood sorting and segregation, and even public investment in schooling (Bell, 2009; Harris & Larsen, 2015; Maroulis et al., 2016; Schneider & Buckley, 2002). Additionally, the policy relevance of this question extends to educational agencies adopting and revising teacher and school accountability frameworks that implicitly attribute school-to-school differences to organizational functioning and seek to carry out instructional improvement efforts in targeted schools (Phillips, Ferguson, Rowley, 2018).

We theorize that three interrelated sets of mechanisms might result in substantial school-to-school differences in instructional quality: (1) student and teacher sorting, (2) organizational functioning, and (3) teacher (mal)adaptation to school context. These mechanisms have received much attention in the literature on educational inequality (e.g., Bryk & Schneider, 2002; Clotfelter, Ladd, & Vigdor, 2006; Rumberger & Palardy, 2005). Yet, even greater concern with school-to-school differences in instructional quality, particularly among educational researchers, has long been suppressed by the well-established finding in the "school effects" literature that measured differences in learning outcomes simply are not that disparate across schools (Scheerens & Bosker, 1997). At the same time, in the U.S. policies to address educational inequality have shifted from the school-focused era of accountability ushered in by the No Child Left Behind Act in 2001 to accountability systems much more focused on the individual teacher

as the lever of change (Close, Amrein-Beardsley, & Collins, 2018). While measures of school quality and accountability currently remain part of the U.S. policy landscape (Current federal education law, ESEA Section 1111(c)(4)(D) requires states to identify and provide comprehensive support and improvement to low performing schools), much debate exists about the relative emphasis that should be placed on school-to-school differences in opportunity to learn.

When studies of student and teacher sorting along with school organizational functioning are paired with the school effects literature then, a paradox emerges: if there are serious sorting mechanisms at work, why are differences in learning gains across schools so small? While we don't expect to fully resolve this paradox, we anticipate that examination of more proximal teaching practice outcomes would reveal larger between school differences than generally reported in studies of school effects, reinvigorating research and school improvement concerns both with student and teacher sorting/segregation and with school organizational features as drivers of inequality in opportunity to learn. Thus, in this study we analyze the magnitude of school-to-school variation in instructional practice relative to the total teacher-level variation in practice, posing two research questions:

1. What is the magnitude of school-to-school variation in instructional practice, as captured by multiple measures?

2. To what extent are school-to-school differences in instruction associated with compositional characteristics of students?

As part of Question One we further explore whether, some domains of instruction are more variable between-schools than other domains. For instance, behavioral management emerges as a common feature across different measures of instruction. In analyzing ratings of

behavioral management from classroom observations as an outcome measure, we label these ratings in aggregate *school behavioral climate*. We acknowledge that theories of instruction and school context treat behavior and instruction as mutually reinforcing, student behavior is both a determinant and outcome of instruction (Metz, 1978; Pace & Hemmings, 2007). However, our measures do not allow us to disentangle that process, and we thus view the measure simply as the enacted behavioral climate during classroom instruction.

Question Two addresses whether the school-level variance in instruction is systematically related to student background, and thus a potential driver of specific gaps in educational outcomes across student groups. As part of Question Two we further explore the potential sources of compositional effects, attempting to isolate basic sorting and selection processes from teacher adaptation. While both sets of processes might be simultaneously ameliorated to some extent by school desegregation and staffing policies, adaptation, particularly maladaptation, might also be addressed directly through instructional reform.

Social Forces Driving School-to-School differences in Opportunity to Learn

Student and teacher sorting. Are students and teachers sorted between schools such that large differences in teaching quality might exist? To date, large-scale evidence exists primarily concerning *teacher* quality, including basic teacher qualifications, rather than the actual enacted quality of instruction. The nationally representative Schools and Staffing Survey (SASS) provides basic information on the distribution of teacher quality in the United States.[1] SASS indicates that the majority of teachers meet *basic* quality standards, even in high-poverty schools. For example, in 2011-2012, 87.8% of teachers in high poverty schools (76-100% poor) compared to 92.6% of teachers in low poverty schools (0-25% poor) had full state-certification,

an important gap but one smaller than might be inferred from popular discussion of "failing" schools (Kelly, Pogodzinski, & Zhang, 2018).

Yet, non-trivial differences in teacher qualifications exist across poor and non-poor schools, including teachers' educational attainment and years of experience, and this is often revealed in state-level administrative data. For example, in Lankford, Loeb, and Wyckoff's (2002) classic study of the teacher labor market in New York State, the relative risk of having a teacher who failed the state's general knowledge exam was approximately 38% higher for the average poor student than the average non-poor student (a probability of .279 vs. .202). Among the state's non-white students, the relative risk of having a teacher who failed the state exam was almost three times higher than among white students (a probability of .212 vs. .071), while the risk of having a teacher with a Bachelor's degree from a least competitive college (as measured by the Barron's ranking of selectivity) was more than twice as high for non-white students (a probability of .214 vs. .102). These latter examples constitute very pronounced differences in basic teacher attributes across some schools and districts.

More recent data from New York show that teacher qualifications in high poverty schools are improving (Lankford et al., 2014), but overall, disparities remain across school, district, and regional boundaries (Adamson & Darling-Hammond, 2012; Goldhaber, Gross, &, Player, 2010; Goldhaber, Lavery, & Theobald, 2015; Schultz, 2014). These disparities are likely to remain, as long as segregation across district boundaries create incentives for the most highly qualified teachers to move to higher socio-economic status schools, where they find more favorable behavioral climates and higher salaries (Kelly, 2004; Guarino, SantiBanez, & Daley, 2006; Ingersoll, 2001).

School organizational functioning. The classic nested layers model of school functioning posits that school resources (e.g., instructional time, learning materials, etc.) flow downward from the district to schools and classrooms to affect learning. Yet, this model has limited power to explain differences in school outcomes, especially when provision of a given resource is at a high enough level where further inputs would yield diminishing returns, or when there simply is not much variation across districts or schools in resource allocation. For example, in Barr & Dreeben's (1983) nested layers model of early literacy development, instructional time-on-task did not vary meaningfully at the district and school level; it was only when teachers began to differentially emphasis literacy at the class and reading group level that differences in opportunity to learn emerged. In contrast, updated nested layers models (Gamoran, Secada, & Marrett, 2000) now conceptualize school resources more broadly to include human and social capital resources, and identify school level organizational attributes as reciprocally influenced by teachers and administrators—such as when teacher collaboration creates shared instructional knowledge and values.

Within this more robust understanding of school organization, multiple processes and domains have been studied in the school improvement literature including: curricular organization (Domina & Saldana, 2012; Klopfenstein, 2004), behavioral climate (Jain et al., 2015; Kelly, 2010), relational trust (Coburn & Russell, 2008; Bryk & Schneider, 2002; Penuel et al., 2010), and academic press (Bryk & Schneider, 2002; Bryk et al., 2010). Here we provide just two examples of how large school-to-school differences can be in these domains. First, behavioral climate, while certainly a socially-constructed concept dependent in many ways on flawed assumptions about student misbehavior (Lewis & Diamond, 2015), appears, from principals' and teachers' own perspective, to differ greatly across schools. In schools with a poor

behavioral climate, principals are three times more likely to report that "disrespect for teachers is a daily problem" than principals in schools with low problem behaviors (Kelly, 2010; see also Jain, 2015). Poor behavioral climates likely affect instruction directly by reducing time on task and engagement (Brophy, 1983; Simonsen et al., 2008), but also indirectly by eroding teacher commitment and effort (Collie et al., 2012). Second, curricular offerings and assignment processes can vary greatly across secondary schools (Kelly & Price, 2011; Bottia et al., 2016; Klopfenstein, 2004; Yun & Moreno, 2006). For example, comparing clusters of schools in California, Yun and Moreno (2006) found that schools serving low-poverty, predominantly Asian and White students offered an average of 11.6 AP courses compared to less than 5 in schools serving high-poverty Latino and Black students.

While we do not measure and disentangle the effects of various school organizational features in this analysis, we theorize that organizational functioning is an important underlying mechanism creating school-to-school differences in instructional practices. Specifically, we hypothesize that measured differences in behavioral climate in particular at the school level are likely related in important ways to school organizational features.

Collective teacher adaptation. In addition to selection processes and school level organizational features, teachers may adapt instruction in response to the composition of their school and classrooms, potentially increasing between-school differences in instruction. Both macro- and micro-adaptations of curriculum and pedagogy to meet student needs is conceptualized as a core tenant of effective instruction (Corno & Snow, 1986; Corno, 2008; Parsons et al., 2018). Beginning teachers encounter the "injunction to adapt" as a central feature of teacher education curricula, including tailoring teaching methods for culturally diverse students but also the broad need to match students' developmental needs (Everitt, 2012). In

reading instruction for example, the need to align instruction with the specific areas of literacy with which the student struggles (Wonder-McDowell, Reutzel, & Smith, 2011), and to use texts that provide appropriate challenge for a given reading level (Allington, 2011) are central tenants of instructional practice.

However, some research has documented *maladaptation* to student background and classroom composition, including both over-adjustment of curriculum and negative responses to challenging behavior or interpersonal dynamics. In a meta-analysis of the quality of teacher-child relationships, Nurmi (2012) concluded that low levels of student motivation and engagement evoke conflict and reduce teacher-child closeness. At the classroom and school level, persistent misbehavior may cause teachers to adapt their instruction to emphasize behavioral management rather than authentic student engagement. Even when overall patterns of adaptation are generally positive, as in reading instruction, maladaptation in specific contexts may arise. For example, focusing on differences by student track level in 8th grade, Northrop & Kelly (2019) find that on a variety of dimensions, including the complexity of the texts teachers select for students, instruction is more disparate than would be predicted based on student achievement level. This finding is consistent with literature on teacher adaptations in expectations and evaluative standards (Kelly & Carbonaro, 2012; Stevens & Van Houtte, 2011). Thus, while the best teachers find ways to modify instruction for individual students and classes that raises average levels of achievement while reducing the dispersion in achievement, observed patterns of adaptation may be negative in some contexts. As part of our analysis, we provide estimates of the extent to which school-to-school differences in instructional practices might be generated by teacher adaptation to different student characteristics, although those estimates are potentially confounded with measurement bias.

A conceptual model of school-level variation in instructional practices. Figure 1 summarizes our conceptual model of school-to-school variation in instructional practices. This model is heavily influenced by Gamoran et al.'s (2000) updated nested layers model, and as such, a two-way arrow depicts the reciprocal relationship between instruction and organizational functioning. We choose to depict collective teacher adaptation as a separate construct from school organizational functioning more generally in order to suggest that adaptation might operate primarily as an aggregate rather than compositional construct (i.e. the aggregate effect of teachers responding individually without reference to spill-over effects from teacher to teacher) In contrast, school organizational functioning in the updated nested layers model is treated as more genuinely compositional in nature, as a set of emergent phenomena with surplus consequences beyond any one individual teacher's actions or intent. Further, no valence is attached to "adaptation" in the figure, but we posit that maladaptation may sometimes occur which exacerbates differences in opportunity to learn across schools.

Note that this model is basic and limited in several ways. First, dynamic processes are not clearly specified. Student and teacher sorting would ostensibly be most a-priori in this model, as long-standing attendance policies and residential patterns create the conditions under which schools operate. Yet, the double-headed arrows on the left of the model are used to show that school organizational functioning and collective teacher adaptation might in turn further shape sorting processes in a feedback loop. Second, we make no judgements about the relative strengths of different effects/paths, although that might be possible with further evidence.

School-to-School Differences in Achievement and Achievement Growth

Point-in-time achievement. Despite the compelling conceptual model offered by the updated nested layers model and empirical studies supporting that model, analyses of variation in

achievement across schools, and especially variation in achievement growth, find quite small school-level variation compared to much larger differences among students within schools. One of the most surprising findings from Coleman and colleagues' 1965 *Equality of Educational Opportunity* report was that the overwhelming majority of the variance in student achievement lies within rather than between schools, and that most of the between-school variance that is apparent likely reflects aggregate compositional effects of students rather than direct effects of school resources or other organizational features. The specific data Coleman analyzed has since been re-examined using modern statistical techniques, and more salient school-to-school differences were revealed (Konstantopoulos & Borman, 2011). However, Coleman's overall insights remain robust. Scheerens and Bosker (1997) carried out a meta-analysis of 168 studies where the gross effect of school-to-school differences could be estimated. While noting that conclusions about school effects depend in part on how between-school variation is expressed/summarized, as well as assumptions about measurement error, Scheerens and Bosker (1997) estimate that perhaps only 9% of the variance in achievement lies between schools (p. 79), and that Coleman's basic conclusions about the size of school effects and the prominent influence of student background were correct (p. 300).

Achievement growth. More recently, educational researchers have specifically analyzed the identification of high- and low-performing schools using student test score gains as opposed to cross-sectional achievement scores. Schochet and Chiang (2010) analyzed 25 subject-specific estimates identifying the proportion of variance in student test scores gains at the school-level (also called the Intra-Class Correlation Coefficient or ICC) from several studies (e.g. Nye, Konstantopoulos, & Hedges, 2004). Across studies, they report an average school-level ICC of .011, indicating very little of the variance in test-score gains is located between schools. Stated

differently, it seems that the vast majority of schools would best be described as "about average." Certainly, any inclination to label large numbers of schools as "failing" (as was the case in some U.S. states during the No Child Left Behind era of federal policy) should be held in check.

The seasonal growth framework. Research that relies on seasonal comparison designs to compare learning trajectories and other outcomes during the school year to the summer, when schools are not in session and the influence of social context is more pronounced, supports and extends basic school effects research. Focusing on gaps between sociodemographic groups, seasonal comparison studies find that schooling is generally compensatory, socioeconomic gaps in achievement outcomes grow much faster when school is not in session, even in the context of high levels of school segregation (Downey, von Hippel, & Broh, 2004; Entwisle, Alexander, & Olson, 1992).[2] Similar findings have been found with non-cognitive outcomes (Downey, Workman, & von Hippel, 2019) and even health outcomes (von Hippel et al., 2007).

School-level variation revisited. Returning to basic questions about variation in school quality, the present analysis is motivated by the concern that even if most schools are "close to average," and/or can be characterized as principally compensatory, analyses of student achievement outcomes may understate or fail to reveal important school-to-school differences that affect the quality of classroom experiences. For instance, this could be because they are not comprehensively capturing all the useful skills that students are learning in schools. Analyzing state data from Texas and Massachusetts, Jennings et al. (2015) find that school effects on the probability of attending a four-year college are somewhat larger than effects on test score gains.[3] Likewise, analyses of the recent Measures of Effective Teaching (MET) Study, the same data we analyze in the present study, suggest that at the teacher level, direct measures of instruction, student experience, and teacher knowledge and preparation provide additional insight into future

achievement growth than current test score gains alone (Kane et al., 2012). At the same time that multiple measures helped to identify underlying teacher effectiveness that is stable across years in the MET data, observations and student reports also captured dimensions of effective teaching not well-captured by test scores (Mihaly et al., 2013).

Building on this literature, the present study is motivated by the assumption that instructional processes at the school level constitute an important dimension of school effectiveness; analogous to teacher-level results, we believe school-level instructional quality says something real about opportunity to learn at that school. Moreover, like Jennings et al. (2015) we anticipate that school-level instructional practices affect non-cognitive outcomes and long-range academic outcomes not captured by a given year's test scores (see Jennings et al., 2015 for further discussion). In all then, we worry that an understanding of how school-to-school differences in opportunity to learn affect educational inequality based on measured achievement alone is incomplete.

Prior research on School-to-School Variation in Instruction: Limited Evidence from Survey Reports

Many studies examine relationships between instructional practices and achievement growth including (e.g., McCaffrey et al., 2001; Newmann et al., 1996; Von Secker & Lissitz, 1999). In contrast, to date there have been few large-scale studies of instructional practice that treat instruction as the variable of interest and seek to explain the sources of school-to-school variation in instruction. The studies we review here find that differences in instructional practice are smaller than differences in organizational features such as curricular organization, etc.

However, at the outset, it should be noted that this inference is limited by the available measures of instruction, which rely heavily on student and teacher survey reports.

Several studies have examined school-to-school variation in English and language arts instructional practice (Kelly & Majerus, 2011; Gamoran & Carbonaro, 2002; Raudenbush et al., 1993). The studies generally find that 5-25% of the total variance in instructional practices lies between schools. For example, Kelly and Majerus (2011) examined teacher-reports of instructional practices in the Chicago School Study data, using Newmann, Marks, and Gamoran's (1996) Disciplined Inquiry framework, finding ICCs ranging from .04 to .09 on dimensions of high-quality ELA instruction.

Raudenbush et al. (1993) investigated teachers' emphasis on higher order thinking skills in mathematics and science instruction, finding less than 1% of the variance in math instruction between-schools and 1.4% of the variance between schools in science. Kelly (2010) studied the use of developmental/student-centered practices in mathematics instruction in Public and Catholic schools. Although only 6% of the teacher-level variance occurred at the school-level, he did find sector differences in use of student-centered teaching practices. Similarly, Northrop and Kelly (2018) find limited variation in developmental instruction in Math and Science between schools, ICCs of .084 and .110 respectively in the High School Longitudinal Study data.

Recently, Phillips, Ferguson, and Rowley (in progress) have analyzed school-to-school variation in student reports of the Tripod 7Cs instructional dimensions (relative to the total teacher-level variation). The authors find 11.5% of the variance on the 7Cs composite score at the middle-school level and 11.6% at the high school level.

In addition to instructional features such as higher-order thinking, coherence, rich assessments, etc., school behavioral climate is an underlying organizational feature of classrooms

and schools affecting opportunity to learn. In 2011-2012, 40.7% of public school teachers somewhat or strongly agreed that student misbehavior interfered with their teaching (Robers et al., 2015), and as previously noted, this is highly variable at the school level. In particular, several studies have documented the relationship between school socioeconomic composition and behavioral climate (Jain et al., 2015; Voight, Geller, & Nation, 2014; Wildhagen, 2012).

In this study we contribute to the literature on school instructional practices by examining an especially broad range of instructional practices. We examine observational measures scored with well-developed protocols by trained raters rather than relying on teacher and student reports. Teacher reports in particular may be biased toward over-reporting use of well-accepted practices.

Data and Methods

The Measures of Effective Teaching Study collected data on teachers' instructional practices in six U.S. school districts over a two-year period from 2009-2010 to 2010-2011. We summarize pertinent features of MET here (The MET user guide, ICPSR document 34771, provides full study details). Of critical importance in pursuing this analysis of the MET data was the simultaneous availability of so many high quality measures of instruction, including: the Danielson Framework for Teaching or FFT (The Danielson Group, 2011), the Classroom Assessment Scoring System or CLASS (Hamre et al., 2013), a condensed version of the Mathematical Quality of Instruction or MQI lite (Hill et al., 2008), the Protocol for Language Arts Teaching Observation or PLATO (Grossman et al., 2014), the Tripod Student Perception survey (Ferguson & Danielson, 2014), as well as assessments of teacher content knowledge for teaching specific to grade level and subject. While these measures are frequently used for teacher evaluation and other applied applications in schools, there is growing interest in the

United States and internationally (Dewulf, 2019; van de Grift, 2007) in using these and other instructional measures in research on school organization and policy.

A second basic feature of MET was collecting two years of data from the same teachers. The present study utilizes both the Year 1 and Year 2 data, appending/stacking the data to increase sample size. This feature of the data, along with a substantial percentage of teachers with more than one class in Year 1 (53%), necessitates using a three-level model (School, Teacher, Class) to estimate the proportion of between school variance. Throughout the analysis, we are focused on the proportion of variance in teacher-level instructional practices that lie between schools. In this approach, year-to-year variance appears in the denominator of our decomposition of variance, such that the between-school variance is the stable school effect across two years of data. However, preliminary analyses show those estimates are very similar to single-year estimates. Investigating only Year 1 data, when teachers were naturally matched with classrooms, we find generally similar ICC estimates (but with larger standard errors). Investigating both years of data, but adding a control for year (affecting the denominator), increases the proportion of between-school variance by as little as only .01.

We focus on grades 4-8, examining school-to-school differences in schools serving grades 4-5, and 6-8 respectively. We chose these groupings as schools serving grades 6-8 are typically termed "middle schools" and these groupings reflect the modal grade structure in the MET sample. In Year 1, 2,741 teachers from 317 schools participated in MET, contributing 3,213 class sections of data in grades 4-8 (English and language arts, math, or for some elementary classrooms, combined ELA and math sections). Raw sample sizes vary substantially across outcome measures, both because some observational protocols pertain to both English and math (FFT, CLASS), while others are subject specific (e.g., PLATO), and because some

observational protocols were utilized to code a smaller subset of lessons/sections. Because of our interest in contrasting school-to-school differences across different observation protocols, and making inferences from fixed-effects models (see below), we want to ensure these comparisons draw on a common sample of teachers within each school, with two years of available data (and multiple classrooms per teacher). Thus, we restrict to the sample of teachers who have FFT and CLASS scores in both years, and for at least one of the years: one or more of the subject-specific measures (MQI or PLATO), CKT scores in at least one subject, and scores on Tripod. This yields a final analytic sample of 2,258 class sections (1,364 in Year 1), taught by 893 teachers in 188 schools from 6 districts.[4]

Within each of the six MET districts, "traditional" public elementary and middle schools were recruited; alternative schools, vocational schools, special education schools, and small schools with fewer than three teachers per grade/subject combination were excluded (this last criteria precluded many charter schools from participating). Teachers were recruited to voluntarily participate in the study, with restrictions including that they were not team teaching/looping, planned to remain at the school for the following year, and were part of an eligible group of teachers that could be randomized in Year 2. Teachers received a $1500 incentive. The MET study included several very large districts, such that the number of teachers and schools participating represents only a very small fraction of the total sampling frame. However, the mean characteristics of the teachers who volunteered for MET are similar to non-participating teachers in those districts. For example, 56.8% of MET teachers were white, with a mean experience of 10.3 years, compared to 59.8% and 11.2 years respectively for non-met teachers (Kane et al., 2012).

Observation and Instructional Measurement Process

In Year 1 of the study, lessons were video recorded during the spring semester (February-June), and spread out in an effort to increase representativeness. Each teacher provided an average of 2-3 videos for each section (e.g., a mean of 2.7 videos were scored on the FFT), balanced between "focal lessons" requested by the MET researchers and lessons of the teacher's choice. Teachers were trained to operate the video and audio recording equipment, which consisted of a camera focused on the board and one providing a 360-view of the room (excluding non-participating students), and two microphones, one for teacher audio and one for overall classroom audio. These were later combined into single channels for lesson scoring.

The observation rating process included 902 current and former teachers using an online platform to score video observations (in addition to the MET user guide, see the MET Observations Measure Report, ICPSR 34771). Videos were scored in four-hour shifts, where raters used a single protocol to score the first 30-35 minutes of each video, often divided into smaller segments of time for a given protocol (the CLASS protocol uses 15-minute segments). Raters were trained over a 17-25 hour period, using a combination of MET developed websites and existing ones associated with a given protocol. Rating quality was further enhanced with calibration videos at the beginning of each rating session, by interspersing "validity videos" into each rater's workload, and by consultation with scoring leaders who "back-scored" a sub-sample of videos to identify raters who needed additional training. Despite generally adequate training and quality control for large-scale research purposes, several measurement limitations have been identified in the MET study, including imprecise discrimination in lesson quality, lack of independence in sub-domains, and sensitivity to rater training (see Kelly et al., 2020 for full discussion of limitations). Yet, at the present time we would still consider the six measures we use state of the art measures of instruction.

Nature of the observational measures. The protocols used in MET are "global" protocols (Kelly et al., 2020), applying overall scores to multiple domains of practice from viewing segments of instruction. For example, CLASS is divided into three broad domains, emotional support, classroom organization, and instructional support, with those in turn comprised of more specific sub-domains (e.g., quality of feedback). Second, the protocols adopt a formative perspective on teaching quality, which is essential to their use in providing teachers feedback. That is, separate domains of instruction collectively constitute effective instruction but have no necessary underlying covariance. Third, although each protocol has perhaps their own exemplary emphasis (e.g., CLASS has an especially well elaborated emphasis on emotional support), they share a focus on the ways in which teachers challenge, support, and engage students. One limitation of the observational scores in MET is that at least some of them are sensitive to compositional features of the classroom. Analyzing the FFT scores, Steinberg and Garrett (2016) report an influence of classroom achievement composition, while Campbell and Ronfeldt (2018) report additional sociodemographic influences. We address this issue analytically in our analysis, but note at the outset that even unadjusted differences in scores still capture real differences in opportunity to learn due to adaptation and peer influence.

Socio-Demographic Composition and Teacher Background

Measures of student composition at the classroom and school level include: Percent black, Hispanic, Asian, and Other; percent special education; percent gifted; percent English language learner; percent free-lunch; mean student age; mean ELA pre-test score (the state-administered English and language arts assessment for 2009); and mean math pre-test score (state assessments from 2009). The state tests were multiple-choice tests, but we prefer them to the alternative assessments administered in MET for the purposes of this analysis because they

pertain to all students, not just eligible students. Moreover, because the state tests were used for administrative purposes, they likely elicited higher, more even levels of effort than in the MET-administered tests. Among student composition measures, missing data arises from one district that did not report gifted status and another that did not report free-lunch status (Common Core of Data identifiers are not available).

Methods

To examine the extent of school-to-school variation in instructional practice, we conducted a decomposition of variance analysis using summary scores on the four observational protocols, Tripod, and math and English content knowledge. Results are presented separately for elementary and middle school. The decomposition of variance was implemented using STATA's 'mixed' command, where maximum likelihood estimates of the Intraclass Correlation Coefficient (in this case, the proportion of variance between schools) can be calculated, in STATA notation, as the ratio of the (School-Level Var(_cons))/(School-Level Var(_cons)+ Teacher-Level Var(_cons)+Var(_Residual)).[5] Additionally, 95% confidence intervals are estimated for each ICC. Tests of statistical significance are then conducted for pair-wise differences in the ICCS with Wald tests using boot-strapped standard errors of the difference. Although there are numerous possible pair-wise tests, generally, we are interested in two sets of comparisons. First, are there statistically significant differences among observational measures, and does a particular measure exhibit notably more or less between-school variance? Second, is there a notable ordering in the between-school variation between student reports of instruction (Tripod), the observational measures, and teacher knowledge measures?

We then turn to examining between-school variation on sub-domain scores, focusing on three of the observational measures (FFT, CLASS, and PLATO). Throughout the analysis, we

primarily analyze the observational measures using the original scales as they would be used by states and districts and as reported by MET (Tables 1, 4, 5, 6, & 7). However, in Table 3, we consider sub-domains of instruction. In the case of FFT and CLASS, a factor analysis indicates that the various subdomains can be grouped into two broad, separable components of teaching, a behavioral climate component and an overall general instructional quality ("General Instruction") component.[6] In the case of PLATO, factor analysis supports PLATO's original "Factor" constructions. Specifically, PLATO's factor constructions named Disciplinary Demand of Classroom Talk & Activity, Instructional Scaffolding, and Classroom Environment are equivalent to what we have renamed Challenge and Discourse, Strategy Use, and Behavioral Climate, respectively. Kane, Staiger and colleagues' (2012, p. 12) analysis of five MET measures also identified a principal component of practices (after the general instruction cluster) capturing classroom and time management (e.g., in CLASS *behavior management* and *productivity*, and in FFT *managing classroom procedures*, and *managing student behavior* domains). Note that, motivated by our interest in school-to-school variation, we use the term behavioral climate rather than behavioral management. While the protocols were intended to gauge teachers' behavioral management methods, in many cases the protocols score the attained behavioral climate of the classroom, and behavioral climate rather than management seems more descriptive when aggregated to the school level in particular. For example, in FFT, an indicator of the "Creating and Environment of Respect and Rapport" domain is respectful talk and turn taking among students. While the logic is that teachers create that dynamic through norm-setting, modeling, etc., the actual indicator is of the achieved climate itself.

Next, we examine differences in instruction associated with compositional characteristics of students at both the teacher and school level using multilevel models (using STATAs mixed

command). First, we report the R^2 as a summary measure of explained variance at the teacher and school levels for the overall scores on the four observational protocols, Tripod, and math and English teacher content knowledge. In this analysis we utilize a simple formula for R^2: ((unconditional variance−conditional variance)/unconditional variance). Second, we examine the explained variance in behavioral climate as distinct from general instruction. Third, we illustrate the school-to-school differences associated with specific student background variables (averaged at the class-level). We consider the reduced-form relationship (removing student-age effects as a nuisance variable) between compositional measures of achievement (variance jointly explained by math and reading achievement), poverty (% free-reduced lunch), and racial composition (variance jointly explained by % black and % Hispanic).

Using teacher fixed effect models to account for non-random measurement error/bias and adaptation to racial composition. Finally, classroom composition can be correlated with instructional practices due to either teacher sorting or an association between classroom composition and teaching practice. To further complicate matters, associations between classroom composition and teaching practice can be due either to measurement bias or to adaptation. To motivate the importance of separating teacher sorting from adaptation/measurement bias, suppose that we observe a strong correlation between racial composition and our instructional practice measures. The optimal policy response depends on, for example, whether low-SES schools have difficulty attracting and retaining high quality teachers, or because teachers adjust their instruction to student learning needs in suboptimal ways. A third option is that in the set of instructional measures used here, raters attribute characteristics of the class to the teacher, a *nonrandom* form of measurement error or bias associated with student composition (Campbell and Ronfeldt, 2018). Policy makers should also

be interested in whether these state-of-the-art measures of instruction, currently being used in schools to evaluate teachers, are penalizing teachers for the predetermined characteristics of their students, although that is impossible to sort out in these data because adaptation and measurement properties are confounded here.

To remove the effects of non-random measurement error that might contribute to school-to-school differences (and affect our estimate of associations between composition and instruction at the school level) and more narrowly identify the possible association between composition and teaching practice due to school-to-school teacher sorting, we utilize an innovative two-stage statistical adjustment strategy relying on variation in instructional practices within teachers but across different classrooms. This approach, which builds on prior teacher fixed effects models (e.g., Burke & Sass, 2013) is described in the online Appendix. While fixed effects models are commonly used to produce purer/more robust estimates of the effects of student characteristics using only within-teacher variance, we have not seen those inferences then used to produce adjusted ICCs that take into account the variance in student characteristics occurring at all levels.

Results

Table 1 reports summary statistics for the variety of measures of instructional practice analyzed here separately for the elementary and secondary grades. We show both overall scores and behavioral climate as opposed to general instruction of relevant protocols. In general, these protocols result in a top-heavy distribution of teacher effectiveness, depicting the typical teacher as effective and the distribution as left skewed. The protocols are also somewhat peakier than a normal distribution; the standardized kurtosis (not reported in Table 1) are between 3 and 4 on average, and even higher in a few cases (e.g., PLATO's behavioral climate has a kurtosis of 10.3

and 5.8 at the elementary and middle level respectively). In all cases we find substantial between school variability, and we turn next to a formal decomposition of variance.

Basic decompositions of variance showing the proportion of variance in instructional practice at the school level in the MET data are reported in Table 2. Level-1 in these data is the instructional session rather than the teacher as in prior research using teacher surveys. At the elementary level, point estimates for the ICC range from a low of .118 to .133 for MQI and PLATO to .402 for English Teacher Knowledge. Among observational measures, CLASS and FFT have the largest ICCs. The ICC for FFT is greater than CLASS (and the difference is statistically significant at the 10% level) in elementary school, yet this is reversed in middle school. Yet, the most salient differences in ICC at the elementary level is the somewhat greater, and statistically significant differences in school-to-school variability in Teacher knowledge measures (.378 and .402 for math and English) relative to observational or student-report measures. At the elementary level, all pairwise tests for differences between teacher knowledge measures and other instructional measures are statistically significant except for the comparison with FFT.

At the middle school level, the ICCs are generally similar. ICCs are smaller at the middle school than elementary level for FFT, Tripod, and both knowledge measures, and larger for CLASS, MQI, and PLATO. The similar or somewhat smaller ICCs at the middle school level possibly reflect that smaller elementary schools (which often enroll students from more homogenous neighborhood catchments), feed into more heterogeneous middle schools (which is supported by subsequent findings on compositional effects).

The results in Table 2 indicate that understandings of between-school differences in instruction depend to a substantial degree on the specific tools and system used to measure

instruction. Additionally, the inferences made from this particular study are substantially affected by sampling error, reflected in relatively large confidence intervals. Yet, overall, the ICCs for the observational measures and student reports in Table 2, which average about .168, are at the upper-end of what is observed for a simple decomposition of variance in student achievement in cross-sectional data, and much higher than reported for gain-scores in achievement. The ICCs for the teacher knowledge measures are considerably higher, as much as twice even upper bound cross-sectional achievement ICCs in some cases. In all cases, the school-to-school variation in the instructional measures considered here greatly exceeds the school-to-school variability seen in achievement gain scores (Schochet & Chiang, 2010).[7]

Table 3 examines the distribution of instructional practices relying on sub-domain information to identify observed behavioral climate as distinct from general instructional quality (capturing the ways teachers challenge, support, and engage students). At both the elementary and secondary level school-to-school differences in behavioral climate are consistently more substantial than that of general instructional quality (and at the elementary level these differences are statistically significant at $p < .05$).[8]

As a frame of reference for the results in Tables 2 and 3, we conducted ancillary analyses to identify the proportion of school-level variance in achievement in the MET data (tabular results not shown). For the district administered math and English tests, which provide estimates most comparable to prior research,[9] we found ICCs for cross-sectional achievement ranging from.169 (elementary ELA) to .218 (Middle school math). For achievement gain scores we considered both the simple gain score model as well as regressor-variable models (see Allison, 1990 for discussion of modeling implications), finding ICCs ranging from .017 (simple change score in elementary ELA) to .086 (regressor-variable estimates for middle school math gains).

These estimates are towards the upper-bound range from prior research, possibly reflecting larger school-level variation in achievement among urban schools, which are prevalent in the MET data compared to a national sample. Yet, the estimates for instructional variation in Tables 2 and 3 are in some cases larger still (and certainly much larger than variation in achievement gains).

The Association between School Composition and Instructional Quality

Table 4 reports variance explained statistics from multilevel models specifying instructional practice as a function of basic student characteristics simultaneously at the teacher and school level. At the teacher level (within-school), the poverty, achievement, and race-ethnic composition of the classroom explain a very limited amount of the variance in instructional practice, 0-5% for these measures. This is expected, as within-school sorting of students in grades 4-8 is not as pronounced as at the high school level, nor would we expect teacher tracking (the matching of teachers to tracked classes) to be particularly pronounced (Clotfelter, Ladd, & Vigdor, 2006).[10] Note that teachers were randomized to students in Year 2 of MET, which can only exert a downward bias on the proportion of explained variance at the between-teacher level. Yet, in practice this has only a minimal effect on the decomposition of variance analysis in Table 4 (see previous discussion in methods section).

In contrast, at the school level, basic compositional characteristics of students explain from 33.1% to 59.2% of the school-level variance in instructional practice at the elementary level, and 18.7% to 45.5% of the variance at the middle school level. Appendix Table A1 reports the underlying coefficients from Table 4 at the Elementary level as an example, revealing the strong effects of racial composition on the four instructional protocols (CLASS, FFT, MQI, and PLATO). At both the elementary and middle level, compositional features explain the least

variance in Tripod student reports and while schools with a greater percent of male students have lower average Tripod scores, predominantly Black and Hispanic schools have higher Tripod scores. Additionally, while the explained variance statistics in Table 4 look substantial for the teacher knowledge measures, the effects of specific aspects of school composition are not so clear in these data, with relatively large standard errors given the size of the coefficients.

Summarizing across the observational measures of instruction in Table 4, about 30-60% of the variance in instructional practice at the elementary level, and 20-45% at the middle school level is explained by these basic covariates. The greater explained variance at the elementary level is a function both of greater variance in composition itself, but also of a stronger underlying relationship between composition and instruction. Thus, it appears that basic student and teacher sorting process are associated with important school-to-school differences in instruction, especially at the elementary level. In ancillary models we also examined explained variance statistics separately for behavioral climate and general instruction, and we find that generally a greater proportion of the school-level variance in general instructional practice is explained.

Appendix Table A2 provides a more explicit analysis of the sources of school-to-school variation in instruction, reporting semi-partial explained variance estimates that account for the correlation of a given classroom composition measure with all other observed measures of classroom composition. For instance, we present explained variance by average math and reading achievement after controlling for pct. male, pct. SPED, pct. ELL, pct. qualifying for free and reduced-price lunch, pct. Hispanic, pct. Black and average age. These results show that of the factors considered in Table 4, racial composition is the most important socio-demographic characteristic for explaining between school variance. Appendix Table A3 reanalyzes the effect of these specific compositional variables separately by behavioral climate and general instruction

outcomes, finding the associations with composition are somewhat stronger for general instructional practice than for behavioral climate.

Reassessing school-level differences in instruction associated with race/ethnicity: accounting for nonrandom measurement error and teacher adaptation to student composition. Table 5 reports coefficients from teacher fixed effects models of instruction examining only the within-teacher variation in instruction between Year 1 and 2 in order to isolate adaptation/measurement effects of classroom composition from matching of students to teachers based on fixed unobservable differences. How sensitive might the measures used here be simply to sociodemographic characteristics of students? Or, focusing on an alternate explanation for the same findings, might teachers adapt or change instruction in response to student composition, potentially in negative ways? Columns 1 and 3 report the reduced-from, within-teacher association between classroom racial composition and instruction in elementary school as measured by CLASS and FFT. We see that race is a statistically significant predictor of FFT in middle school and CLASS in elementary school—when the same teacher has a higher share of black and Hispanic students in their classroom, the teacher receives a lower rating. Columns 2 and 4 include the full set of measured covariates. Percent male and percent black have statistically significant partial associations with instruction as measured by CLASS, but collectively, all covariates explain only 4.5% of the within-teacher variance in CLASS and 0.5% for FFT in elementary school (1.3% of the within-teacher variance in CLASS and 3% for FFT in middle school.) We interpret these findings as possible evidence of a minor level of teacher maladaptation to student composition, but at the same time, if these relationships reflect measurement bias (i.e. a direct effect of student composition on the measures themselves, independent of teacher behavior), then the measurement bias is relatively minimal. Estimates

from these models and the related models for the other dependent variables are next used in Table 6 to consider school-to-school variation accounting for any measurement bias and adaptation due to student composition.

Table 6 examines the variance in instruction explained at the school level by student composition covariates, after statistical procedures to more narrowly capture the effect of teacher selection (or the differential sorting of teachers to schools with different student body compositions). Whereas the explained variance in Table 4 reflects an unknown combination of selection, other school level organizational features (e.g., school leadership), adaptation, and measurement bias, the latter two effects have been removed from Table 6. A substantial association between student composition and instructional practices remains, explaining from 18.9% (PLATO) to 58.3% (MQI) of the variance at the between-school level for elementary school, and 12.3% (FFT) to 62.8% (MQI) at the middle school level. Although the evidence is indirect (we are not observing teacher moves), we believe this reveals a likely role of teacher sorting in generating school-to-school differences in instruction. However, we cannot separate teacher selection from other unmeasured school organizational features.

The great variability in the results in Table 6 across the different instructional measures suggests the results should be interpreted with caution. The statistical adjustment to isolate and remove measurement bias and adaptation effects is a very rough-hewn inferential process. The adjustment is only as good as the quality of the information from the within-teacher variance. Given the overall level of measurement error in the data, it would be possible to *over-adjust* teachers' practice (which is perhaps reflected in the lower bound estimates in Table 6). As a whole though, the results in Table 6 seem to support our concern that the associations reported in Table 4 are due in part to selection, but also a larger mix of factors including adaptation.

DISCUSSION

In combining a large-scale video-based recording of instructional practice with an extensive coding analysis using multiple protocols, the MET study gives us a view behind classroom doors seldom seen by most parents, or indeed even most educators. Kelly and Northrop (2013), informed by school effects research, argue that the public has a naïve view of school-to-school differences in instruction, believing that instructional practices are much more variable between different schools than they really are. Likewise, Downey and Condron (2016) have argued that the compensatory nature of schools is not fully appreciated by social science researchers, such that other policy levers that might affect child development and achievement are overlooked. These views contrast with educational research in the critical theory paradigm, which argues that pronounced differences in opportunity to learn between schools is an important driver of educational inequality (Carter, 2016; Darling-Hammond, 2010; Demereth, 2009).

The basic descriptive results from this study suggest a middle-view of school-to-school differences in instruction. We find that: first, substantial school level variation in instruction exists, with 30% or more of the total variance in instruction lying between schools in these data. Second, the behavioral climate during instruction appears to be particularly salient, and especially in elementary schools. Third, much of the between school variance, in some cases 40% or more, is readily explained by simple measures of socio-demographic composition, including in particular the racial make-up of schools in the MET districts. Fourth, some evidence from a statistical adjustment method suggests that teacher sorting, rather than measurement bias and teacher adaptation is principally responsible for school-to-school

differences in instruction, although this method will need to be further interrogated in future studies.

The nature of the measures employed here allowed us to investigate school-to-school differences in multiple ways. FFT and CLASS are especially comprehensive measures of instructional activities and processes, while PLATO and MQI have the advantage of being tailored to specific subjects, and thus, somewhat more fine-grained. Tripod has the advantage of pooling data from approximately 20 students per class, leading to reliable estimates of student perceptions of teacher-student interaction quality. The teacher knowledge measures, while not actual measures of instruction, are closely conceptually related to pedagogical decision-making (Clarke & Hollingsworth, 2002; Hill et al., 2008). We were also able to consider behavioral climate as distinct from general instructional quality. On several protocols, we found larger variability in behavioral climate than that of general instruction. An important consideration in interpreting this finding is that among the protocol subdomains, scores on behavioral climate were generally highest; e.g. for CLASS, 85% of lessons scored at the "high" end of the distribution (Kane et al., 2012). Thus, the variation observed in these data is largely because some schools have excellent behavioral climates while others are just "good." Moreover, it is general instruction that is most strongly associated with school composition, rather than behavioral climate. We make the post-hoc hypothesis that school leadership efforts in many urban schools are especially attentive to student behavior, and as a result, have substantially mitigated routine problems of student behavior that affect classroom instruction. However, the nature of the measures is a caveat here; these measures pick up what might be called "routine" problem behaviors seen in class. School-to-school variability in more serious student behavior offenses may differ.

There are however, a number of limitations to the view of instruction captured in the MET study. First, the MET data are a convenience sample of districts, are not necessarily representative, and thus do not necessarily capture the average student's choice set. However, the basic decomposition of variance in achievement outcomes aligns well with prior research. Second, an important feature of many of the MET observational protocols, including the two protocols applicable to multiple subjects, FFT and CLASS, is that they do not solely capture the teacher's own contribution to instruction apart from adaptation and response to students. For example, consider two indicators of the FFT sub-domain, *Creating an Environment of Respect and Rapport*. Rubric example indicators for the proficient category include, "teacher greets students by name as they enter the class or during the lesson" and "students attend fully to what the teacher is saying." The first example is more narrowly a "teacher move," while the second seems influenced by what students bring to the class. This feature of the observational protocols makes it difficult to estimate how the uneven sorting of teachers alone, apart from student composition, might contribute to school-to-school variation in instruction. We made an analytic effort to quantify the extent of measurement bias and adaptation to race/ethnic composition using within-teacher fixed effect models, and we believe this may be an important component of school-to-school differences, along with teacher sorting.

A third limitation is that while MET provides an impressively comprehensive portrait of instruction, as previously discussed, there is a great deal of measurement error in these observational protocols. Ho and Kane (2013), reporting on the MET study, concluded, "…Researchers and school districts should continue to refine their observational instruments to provide even richer, more discerning feedback to teachers."

Within the context of these limitations, the MET data reveal important school-to-school differences in instruction, and that these differences are associated with the student and teacher composition of schools. What policy implications might follow from these descriptive results? Keeping in mind the limitations described, we make the initial hypothesis that observed differences in instruction are caused by teacher sorting in response to student composition (see e.g., Goldhaber, Gross, &, Player, 2010). We acknowledge that instruction is likely additionally affected by student composition itself; even if teachers were randomly distributed across schools, differences in student composition might produce variability in school behavioral climate. In discussing policy implications however, it is worth focusing on teacher sorting.

Unfortunately, In the United States, trends in school segregation coupled with current education laws make addressing student composition differences difficult in many regions of the country (Clotfelter, Ladd, & Vigdor, 2006). Thus, it is necessary to find ways to address uneven teacher staffing within a policy context of high levels of student segregation. Teacher sorting might be addressed with selective incentive programs that reward high quality teachers for teaching in hard to staff schools. A number of selective incentive programs have successfully impacted teacher sorting (Clotfelter et al., 2008; Cowan & Goldhaber, 2015; Springer et al., 2009). Null findings have also been reported in several incentive programs (e.g., Glazerman & Seifullah, 2010; Steele, Murnane, & Willett, 2009). Moreover, the positive impacts of incentives may ultimately fail to overcome teacher attrition due to challenging climate and working conditions in high poverty schools. We remain optimistic that selective incentive programs, especially aggressive ones, might be effective if paired with other policy efforts. First, well-designed teacher pipeline programs can be used to increase the overall pool of highly trained and committed teachers entering the workforce (Henry, Bastian, & Smith, 2012). Second, creative

curricular programs might be used to draw both students and teachers to socio-economically diverse schools (see e.g., Olson Beal & Beal, 2016). A three-pronged approach of effective teacher recruitment and socialization, selective incentives, and curricular reform and innovation could help to produce a more even distribution of instructional quality across schools.

RESEARCH ETHICS

The data analyzed in this study were collected under the supervision of an appropriate ethics committee; all human subjects gave their informed consent prior to their participation in the research. Adequate steps were taken to protect participants' confidentiality pursuant to the MET secure data user policies

REFERENCES

Adamson, F., & Darling-hammond, L. (2012). Funding disparities and the inequitable distribution of teachers: Evaluating sources and solutions. *Education Policy Analysis Archives*, *20*, 1–46.

Allington, R. L. (2011). Reading intervention in the middle grades. *Voices from the Middle*, *19*(2), 10–16.

Allison, P. D. (1990). Change scores as dependent variables in regression analyses. *Sociological Methodology*, *20*, 93–114.

Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.

Bell, C. (2009). Geography in parental choice. *American Journal of Education*, *115*, 493–521.

Bottia, M. C., Giersch, J., Mickelson, R. A., Stearns, E., & Moller, S. (2015). Distributive justice antecedants of race and gender disparities in first-year college performance. *Social Justice Research*, *29*, 35–72.

Brophy, J. E. (1983). Classroom organization and management. *The Elementary School Journal*, *83*, 264–285.

Bryk, A. S., Lee, V. E., & Holland, P. B. (1993). *Catholic schools and the common good*. Cambridge, MA: Harvard University Press.

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for improvement*. New York: Russell Sage Foundation.

Burke, M. A., & Sass, T. R. (2013). Classroom peer effects and student achievement. *Journal of Labor Economics, 31*, 51–82.

Campbell, S., L. & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*, 1233–1267.

Carter, P. L. (2016). Carter common on Downey and Condron. Sociology of Education, 89, 225–226.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*, 947–967.

Close, K., Amrein-Beardsley, A., & Collins, C. (2018). *State-Level assessments and teacher evaluation systems after the passage of the Every Student Succeeds Act: Some steps in the right direction*. National Education Policy Center.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *The Journal of Human Resources, 41*, 778–820.

Clotfelter, C. T., Glennie, E. J., Ladd, H. F., & Vigdor, J. L. (2008). Teacher bonuses and teacher retention in low-performing schools: Evidence from North Carolina $1800 teacher bonus program. *Public Finance Review, 36,* 63–87.

Coburn, C.E., & Russell, J.L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis, 30*, 203–235.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfall, F. D., & York, R. L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Department of Health, Education and Welfare.

Collie, R. J., Shapka, J. D., & Perry, N. E. (2012). School climate and social–emotional learning: Predicting teacher stress, job satisfaction, and teaching efficacy. *Journal of Educational Psychology, 104*, 1189–1204.

Connell, R. W., Ashenden, D. J., Kessler, S., & Dowsett, G. W. (1982). *Making the difference: Schools, families, and social division*. St Leonards, AU: Allen & Unwin.

Corno, L. (2008). On teaching adaptively. *Educational Psychologist*, *43*, 161–173.

Corno, L., & Snow, R. E. (1986). Adapting teaching to individual differences in learners. In M. C. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 605–629). Washington, DC: American Educational Research Association.

Cowan, J., & Goldhaber, D. (2015). *Do bonuses affect teacher staffing and student achievement in high-poverty schools? Evidence from an incentive for National Board Certified Teachers in Washington State*. University of Washington Bothell: Center for Education Data & Research.

The Danielson Group. (2011). *The Framework for Teaching Evaluation Instrument*. Princeton, NJ.

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future*. New York: Teachers College Press.

Demereth, P. (2009). *Producing success: The culture of personal advancement in an American high school*. Chicago, IL: University of Chicago Press.

Dewulf, L. (2019). *Teacher beliefs and teaching quality in segregated primary school classes: A study into the effects on students' language achievement*. Ghent University: Dissertation.

Domina, T., & Saldana, J. (2012). Did raising the bar level the playing field? Mathematics curricular intensification and inequality in American high schools, 1982-2004. *American Education Research Journal*, *49*, 685–708.

Downey, D. B., & Condron, D. J. (2016). Fifty years since the Coleman report: Rethinking the relationship between schools and inequality. *Sociology of Education*, *89*, 207–220.

Downey, D. B., von Hippel, P. T., & Broh, B. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review, 69*, 613–35.

Downey, D. B., Workman, J., & von Hippel, P. T. (2019). Socioeconomic, Ethnic, Racial, and Gender Gaps in Children's Social/Behavioral Skills: Do They Grow Faster in School or out? *Sociological Science, 6*(17).

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, Schools, & Inequality*. Boulder, CO: Westview Press.

Everitt, J. G. (2012). Teacher education and accountability: Adapting to prospective work environments in public schools. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement*. (pp. 137–160). New York: Teachers College Press.

Ferguson, R. F., & Danielson, C. (2014). How framework for teaching and Tripod 7Cs evidence distinguish key components of effective teaching. In T. J. Kane, K. A. Kerr, & R. Pianta (Eds.) *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 98–143). San Francisco, CA: Jossey-Bass.

Gamoran, A., Secada, W. G., & Marrett, C. B. (2000). The organizational context of teaching and learning: Changing theoretical perspectives. In M. T. Hallinan (Ed.), *Handbook of the sociology of education* (pp. 37–64). New York: Kluwer.

Glazerman, S., & Seifullah, A. (2010). *An evaluation of the Teacher Advancement Program (TAP) in Chicago: Year two impact report*. Princeton, NJ: Mathematica Policy Research.

Goldhaber, D., Gross, B., & Player, D. (2010). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *Journal of Policy Analysis and Management*, *30*, 57–87.

Goldhaber, D., Lavery, L., & Theobald, R. (2015). Uneven playing field? Assessing the teacher quality gap between advantaged and disadvantaged students. *Educational Researcher*, *44*, 293–307.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*, 293–303.

Guarino, C. M., SantiBanez, L., & Daley, G. A. (2006). Teacher recruitment and retention: A review of the recent empirical literature. *Review of Educational Research*, *76*, 173–208.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L.,… & Hamagami, A. (2013). Teaching through interactions: testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*, 461–487.

Harris, D. N., & Larsen, M. F. (2015). *What schools do families want (and why)? New Orleans families and their school choices before and after Katrina*. New Orleans: Research Alliance for New Orleans.

Harris, D. N., & Larsen, M. F. (2016). *The effects of the New Orleans post-Katrina school reforms on student academic outcomes*. New Orleans: Research Alliance for New Orleans.

Henry, G. T., Bastian, K. C., & Smith, A. A. (2012). Scholarships to recruit the ‘‘best and brightest’’ into teaching: Who is recruited, where do they teach, how effective are they, and how long do they stay? *Educational Researcher*, *41*, 83–92.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study, *Cognition and Instruction*, *26*, 430–511.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. (Tech. Rep.). (Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project).

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*, 499–534.

Jain, S., Cohen, A. K., Huang, K., Hanson, T. L., & Austin, G. (2015). Inequalities in school climate in California. *Journal of Educational Administration*, *53*, 237–261.

Jennings, J. L., Deming, D., Jencks, C., Lopuch, M., & B. E. Schueler. (2015). Do differences in school quality matter more than we thought? New evidence on educational opportunity in the twenty-first century. *Sociology of Education*, *88*, 56–82.

Kane, T., Staiger, D., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., . . . Parker, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Tech. Rep.). (Seattle, WA: Bill & Melinda Gates Foundation, Measures of Effective Teaching Project).

Kelly, S. (2004). An event history analysis of teacher attrition: Salary, teacher tracking, and socially disadvantaged schools. *Journal of Experimental Education*, *72*, 195–220.

Kelly, S. (2010). A crisis of authority in predominantly black schools? *Teachers College Record*, *112*, 1247–1274.

Kelly, S. (2010). The prevalence of developmental instruction in public and Catholic Schools. *Teachers College Record*, *112*, 2405–2440.

Kelly, S., Bringe, R., Aucejo, E., & Fruehwirth, J. (2020). Using global observation protocols to inform research on teaching effectiveness and school improvement: Strengths and emerging limitations. *Education Policy Analysis Archives*, 28(62).

Kelly, S., & Carbonaro, W. (2012). Curriculum tracking and teacher expectations: Evidence from discrepant course taking models. *Social Psychology of Education, 15*, 271–294.

Kelly, S., & Majerus, R. (2011). School-to-school variation in disciplined inquiry. *Urban Education*, *46*, 1553–1583.

Kelly, S., & Northrop, L. (2013). How parents grade schools. *Contexts, 12*(4), 68–70.

Kelly, S., Pogodzinski, B., & Zhang, Y. (2018). Teaching quality. In B. Schneider & G. Saw, (Eds.). *Handbook of the sociology of education in the 21st century* (pp. 275–296). New York: Springer

Kelly, S., & Price, H. (2011). The correlates of tracking policy: Opportunity hoarding, status competition, or a technical-functional explanation? *American Educational Research Journal*, *48*, 560–585.

Klopfenstein, K. (2004). Advanced Placement: Do minorities have equal opportunity? *Economics of Education Review*, *23*, 115–131.

Konstantopoulos, S., & Borman, G. D. (2011). Family background and school effects on student achievement: A multilevel re-analysis of the Coleman data. *Teachers College Record*, *113*, 97–132.

Lankford, H., Loeb, S., Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*, 37–62.

Lankford, H., Loeb, S., McEachin, A., Miller, L. C., & Wyckoff, J. (2014). Who enters teaching? Encouraging evidence that the status of teaching is improving. *Educational Researcher*, *43*, 444–453.

Lewis, A.E., & Diamond, J.B. (2015). *Despite the best intentions: How racial inequality thrives in good schools*. New York: Oxford University Press.

Maroulis, S., Santillano, R., Jabbar, H., & Harris, D. N. (2016). *The push and pull of school performance: Evidence from student mobility in New Orleans*. Working Paper. New Orleans: Research Alliance for New Orleans.

McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Klein, S. P., Bugliari, D., & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education*, *32*, 493–517.

Metz, M. H. (1978). Classrooms and corridors: The crisis of authority in desegregated secondary schools. Berkeley: University of California Press.

Nurmi, J-E. (2012). Students' characteristics and teacher–child relationships in instruction: A meta-analysis. *Educational Research Review*, *7*, 177–197.

Olson Beal, H. K., & Beal, B. D. (2016). Assessing the impact of school-based marketing efforts: A case study of a foreign language immersion program in a school-choice environment. *Peabody Journal of Education*, *91*, 81–99.

Northrop, L., & Kelly, S. (2018). AYP status, urbanicity, and sector: School-to-school variation in instruction. *Urban Education*, *53*, 591–620.

Northrop, L., & Kelly, S. (2019). Who gets to read what? Tracking, instructional practices, and text complexity for middle school struggling readers. *Reading Research Quarterly*, *54*, 339–361.

Nye, B., S. Konstantopoulos, & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26(3), 237–257.

Pace, J. L., & Hemmings, A. (2007). Understanding authority in classrooms: A review of theory, ideology, and research. *Review of Educational Research, 77*, 4–27.

Parsons, S. A., Vaughn, M., Scales, R. Q., Gallagher, M. A., Parsons, A. W., Davis, S. G., Pierczynski, M., & Allen, M. (2018). Teachers' instructional adaptations: A research synthesis. *Review of Educational Research, 88*, 205–242.

Penuel, W.R., Riel, M., Joshi, A., Pearlman, L., Kim, C.M., & Frank, K.A. (2010). The alignment of the informal and formal organizational supports for reform: Implications for improving teaching in schools. *Educational Administration Quarterly, 46*, 57–95.

Phillips, S. F., Ferguson, R. F., & Rowley, J. F. S. *Evaluating the Tripod 7Cs composite as an alternative measure of school quality*. Manuscript in progress.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. (1993). Higher order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal*, *30*, 523–553.

Robers, S., Zhang, A., Morgan, R.E., & Musu-Gillette, L. (2015). *Indicators of School Crime and Safety: 2014* (NCES 2015-072/NCJ 248036). National Center for Education Statistics, U.S. Washington, DC: US Department of Education.

Rumberger, R. W., & Palardy, G. J. (2005). Does segregation still matter? The impact of student composition on academic achievement in high school. *Teachers College Record*, *107*, 1999–2045.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. New York, NY: Elsevier Science.

Schneider, M., & Buckley, J. (2002). What do parents want from schools? Evidence from the internet. *Educational Evaluation and Policy Analysis*, *24*, 133–144.

Schochet, P. Z., & Chiang, H. S. (2010). Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains. NCEE 2010-4004. *National Center for Education Evaluation and Regional Assistance*.

Schultz, L. M. (2014). Inequitable dispersion: Mapping the distribution of highly qualified teachers in St. Louis metropolitan elementary schools. *Education Policy Analysis Archives*, *22*(90), 1–24.

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based practices in classroom management: Considerations for research to practice. *Education and Treatment of Children*, *31*, 351–380.

Springer, M. G., Lewis, J. L., Podgursky, M. J., Ehlert, M. W., Gronberg, T. J., Hamilton, L. S.,…Peng, A. (2009). *Texas Educator Excellence Grant (TEEG) program. Year three evaluation report*. Nashville, TN: National Center on Performance Incentives.

Steele, J. L., Murnane, R. J., & Willett, J. B. (2009). *Do financial incentives help low-performing schools attract and keep academically talented teachers? Evidence from California* (Working Paper No 14780). Cambridge, MA: National Bureau of Economic Research.

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? E*ducational Evaluation and Policy Analysis*, *38*, 293–317.

Stevens, P. A. J., & Van Houtte, M. (2011). Adapting to the system or the student? Exploring teacher adaptations to disadvantaged students in an English and a Belgian secondary school. *Educational Evaluation and Policy Analysis, 33*, 59–75.

Van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational Research, 49*, 127–152.

Voight, A., Geller, J. D., & Nation, M. (2014). Contextualizing the ‘‘behavior gap’’: Student prosocial behavior and racial composition in urban middle schools. *Journal of Early Adolescence, 34*(2), 157–177.

Von Hippel, P. T., Powell, B., & Downey, D. B. (2007). The effect of school on overweight in childhood: Gain in body mass index during the school year and during summer vacation. *American Journal of Public Health, 97*, 696–702.

Von Secker, C. E., & Lissitz, R. W. (1999). Estimating the impact of instructional practices on student achievement in science. *Journal of Research in Science Teaching, 36*, 1110–1126.

Wildhagen, T. (2012). How teachers and schools contribute to racial differences in the realization of academic potential. *Teachers College Record, 114*(7), 1–27.

Wonder-McDowell, C., Reutzel, D.R., & Smith, J.A. (2011). Does instructional alignment matter? Effects on struggling second graders' reading achievement. *The Elementary School Journal, 112*, 259–279.

Yun, J. T., & Moreno, J. F. (2006). College access, K-12 concentrated disadvantaged, and the next 25 years of education research. *Educational Researcher, 35*, 12–19.

Figure 1. A basic conceptual model of school level variation in instructional practices

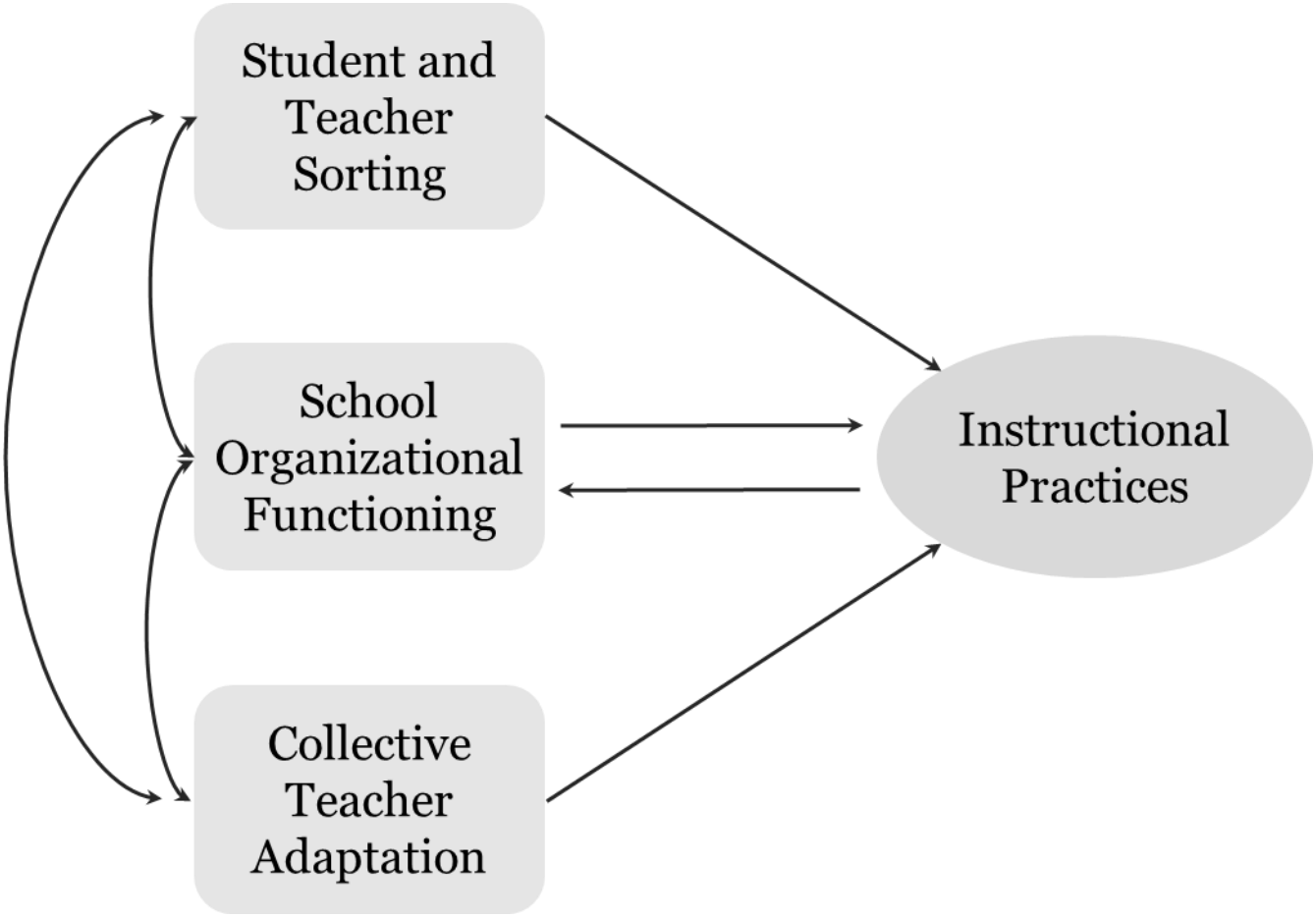

Table 1. Summary statistics: Four observational protocols, Tripod student reports, and teacher content knowledge, elementary (4-5) and middle (6-8) grades.

	Mean	SD	(Min, Max)	SD Between-School	Skewness
Elementary School					
English Teacher Knowledge	52.36	13.97	[20.45, 89.2]	10.73	0.23
Math Teacher Knowledge	66.24	12.1	[33.65, 89.42]	9.18	-0.47
CLASS overall composite	4.55	0.37	[3.05, 5.73]	0.22	-0.2
FFT overall composite	2.66	0.25	[1.69, 3.38]	0.16	-0.34
Overall PLATO score	2.68	0.26	[1.71, 3.54]	0.15	-0.37
Overall MQI score	1.59	0.12	[1, 2.25]	0.07	0.22
Tripod 7 Cs composite	0	0.24	[-1.3, 0.87]	0.14	-0.22
CLASS Behavioral Climate	6.19	0.36	[3.83, 7]	0.25	-1.25
CLASS General Instruction	4.07	0.42	[2.22, 5.67]	0.23	-0.02
FFT Behavioral Climate	2.83	0.26	[1.54, 3.5]	0.18	-1.16
FFT General Instruction	2.55	0.27	[1.75, 3.5]	0.16	0.09
PLATO Behavioral Climate	3.79	0.28	[2.13, 4]	0.19	-2.41
PLATO Challenge and Discourse	2.43	0.34	[1, 3.5]	0.18	-0.18
PLATO Strategy Use	1.81	0.45	[1, 4]	0.25	0.78
Middle School					
English Teacher Knowledge	61.7	14.69	[22.6, 88.94]	10.9	-0.25
Math Teacher Knowledge	64.78	10.78	[35.29, 86.54]	7.83	-0.33
CLASS overall composite	4.14	0.54	[2.42, 5.8]	0.34	-0.21
FFT overall composite	2.43	0.36	[1.06, 3.88]	0.2	-0.32

Overall PLATO score	2.39	0.32	[1.17, 3.58]	0.19	-0.35
Overall MQI score	1.56	0.14	[1.06, 2.38]	0.08	1.15
Tripod 7 Cs composite	0	0.35	[-1.18, 0.99]	0.19	-0.57
CLASS Behavioral Climate	5.98	0.58	[3.17, 7]	0.37	-1.18
CLASS General Instruction	3.57	0.6	[1.58, 5.44]	0.36	-0.02
FFT Behavioral Climate	2.63	0.43	[1, 4]	0.25	-0.86
FFT General Instruction	2.31	0.38	[1, 4]	0.19	0.06
PLATO Behavioral Climate	3.54	0.5	[1.13, 4]	0.28	-1.61
PLATO Challenge and Discourse	2.18	0.43	[1, 4]	0.23	0.18
PLATO Strategy Use	1.46	0.38	[1, 3.25]	0.22	1.22

Table 2. Decomposition of variance in instructional practices within and between schools: Four observational protocols, Tripod student reports, and teacher content knowledge, elementary (4-5) and middle (6-8) grades.

Dependent Variable	Number of classrooms	Proportion of var between Schools (ICC)	95% Confidence interval of ICC
Elementary			
CLASS overall composite	837	.180	.113, .275
FFT overall composite	837	.245	.168, .344
Overall MQI score	645	.118	.064, .209
Overall PLATO score	698	.133	.080, .212

Tripod 7 Cs composite	837	.162	.105, .240
English Teacher Knowledge	656	.402	.324, .496
Mathematics Teacher Knowledge	617	.378	.273, .494
Middle			
CLASS overall composite	1421	.197	.125, .295
FFT overall composite	1421	.138	.090, .206
Overall MQI score	704	.129	.055, .275
Overall PLATO score	723	.198	.135, .282
Tripod 7 Cs composite	1421	.069	.027, .167
English Teacher Knowledge	720	.317	.200, .463
Mathematics Teacher Knowledge	703	.295	.197, .418

Note: At the elementary and middle school level, all pairwise tests for differences between the teacher knowledge measures and other instructional measures are statistically significant at the 5% level, except for the comparison between CKT Math and FFT at the elementary level (p=0.067). FFT is also statistically significantly different than PLATO and MQI at the 5% level. There are 92 elementary schools and 96 middle schools. There are 391 teachers at the elementary level, and 503 teachers at the middle school level.

Table 3. Decomposition of variance in instructional practices within and between schools: Tripod 7 Cs scales and CLASS sub-domains, elementary (4-5) and middle (6-8) grades.

Dependent Variable	Number of classrooms	Proportion of var between schools	95% CI of ICC
Elementary			
CLASS			
Behavioral Climate	837	.298	.234, .370
General Instruction	837	.122	.062, .226
FFT			
Behavioral Climate	837	.268	.188, .366
General Instruction	837	.183	.117, .275
PLATO			
Behavioral Climate	698	.165	.095, .271
Challenge and Discourse	698	.117	.061, .211
Strategy Use	698	.100	.050, .191
Middle			
CLASS			
Behavioral Climate	1421	.188	.116, .289
General Instruction	1421	.179	.108, .282

FFT			
Behavioral Climate	1421	.150	.093, .232
General Instruction	1421	.114	.071, .179
PLATO			
Behavioral Climate	723	.137	.074, .239
Challenge and Discourse	723	.156	.105, .226
Strategy Use	723	.099	.042, .217

Note: At the elementary level, differences are statistically significant comparing behavioral climate and GI within CLASS and within FFT, (p=0 and p=0.030, respectively). This is not the case at the middle school level. There are 92 elementary schools and 96 middle schools. There are 391 teachers at the elementary level, and 503 teachers at the middle school level.

Table 4. Explained variance in instructional practices due to poverty, achievement, and race-ethnic composition within and between schools: Four observational protocols, Tripod student reports, and teacher content knowledge, elementary (4-5) and middle (6-8) grades.

	Proportion of Explained Variance (R^2)		
Dependent Variable	Between-School	Between Teacher, Within-School	Total
Elementary			
CLASS overall composite	.592	.015	.133

FFT overall composite	.545	.017	.135
Overall MQI score	.331	.108	.063
Overall PLATO score	.410	.000[a]	.056
Tripod 7 Cs composite	.323	.089	.076
English Teacher Knowledge	.456	.022	.196
Mathematics Teacher Knowledge	.492	.000[a]	.179
Middle			
CLASS overall composite	.346	.160	.111
FFT overall composite	.455	.084	.097
Overall MQI score	.386	.207	.074
Overall PLATO score	.187	.178	.069
Tripod 7 Cs composite	.149	.032	.039
English Teacher Knowledge	.284	.023	.106
Mathematics Teacher Knowledge	.265	.045	.110

[a] Controls for average initial achievement in math and reading, pct. male, pct. SPED, pct. ELL, pct. qualifying for free and reduced price lunch, pct. Hispanic, pct. Black and average age are included.

[b] Negative values are computationally possible in multilevel explained variance calculations when using the traditional R^2 calculation, and are truncated to zero here.

Table 5. Associations between socio-demographic composition and instructional practices from fixed effects models examining between-year variation within teachers. Standard errors in parentheses

Dependent Variable	CLASS Overall Composite		FFT Overall Composite	
Model	1	2	3	4
Elementary				
Pct. Hispanic	−.88 (.40)**	−.44 (.47)	−.25 (.36)	−.22 (.41)
Pct. Black	−1.55 (.49)***	−1.37 (.53)**	−.34 (.52)	−.30 (.52)
Class-mean math achiev.		−.03 (.24)		.03 (.21)
Class-mean ELA achiev.		.12 (.26)		−.01 (.20)
Pct. Male		−.73 (.37)*		−.39 (.29)
Pct. special education		.14 (.42)		.27 (.36)
Pct. English lang. learner		−.20 (.38)		−.02 (.43)
Class-mean age		−.09 (.07)		.02 (.05)
Pct. Free/reduced lunch		−.62 (.42)		.00 (.31)
R-squared Overall	.110	.119	.128	.127
R-squared Within-Teacher	.022	.045	.001	.005
Middle				
Pct. Hispanic	−.14 (.26)	.24 (.30)	−.61 (.24)**	−.18 (.31)
Pct. Black	−.40 (.21)*	.09 (.26)	−.90 (.21)***	−.42 (.25)*
Class-mean math achiev.		.14 (.14)		−.09 (.14)
Class-mean ELA achiev.		.04 (.15)		.22 (.16)

Pct. Male		−.12 (.18)		−.16 (.19)
Pct. special education		−.09 (.22)		−.41 (.25)
Pct. English lang. learner		.04 (.28)		.13 (.28)
Class-mean age		.02 (.05)		.03 (.05)
Pct. Free/reduced lunch		−.09 (.25)		−.39 (.27)
R-squared Overall	.039	.005	.064	.074
R-squared Within-Teacher	.003	.013	.015	.030

* P < .1, ** p < .05, *** p < .01

Table 6. Explained variance in instructional practices within and between schools, removing variance due to adaptation and measurement error using estimates from within-teacher fixed-effects models (estimation approach further described in the appendix).

	Proportion of Explained Variance (R^2)		
Dependent Variable	Between-School	Between Teacher, Within-School	Total
Elementary			
CLASS overall composite	.367	.032	.052
FFT overall composite	.362	.034	.071
Overall MQI score	.583	.079	.107
Overall PLATO score	.189	.120	.032
Tripod 7 Cs composite	.400	.087	.084

Middle			
CLASS overall composite	.326	.134	.093
FFT overall composite	.123	.080	.029
Overall MQI score	.628	.196	.137
Overall PLATO score	.160	.121	.044
Tripod 7 Cs composite	.170	.027	.025

[a] Negative values are computationally possible in multilevel explained variance calculations and are truncated to zero here for consistency with traditional R^2 calculations (see Snijders & Bosker, 1999).

ENDNOTES

[1] With the 2015-2016 data collection, the National Center for Education Statistics has transitioned SASS to the new National Teacher and Principal Survey (NTPS).

[2] Schooling is compensatory to the extent that it counterbalances or reduces initial deficits. In the Downey et al. analysis, the correlation between initial status and growth during the school year is negative. Arguably, in ratio scale achievement data, even a zero correlation between initial status and growth could be deemed compensatory, as an equal increase in means leads to a real reduction in inequality.

[3] See bottom two rows of Table 4 in Jennings et al. (2015). Note that the ICCs themselves show greater relative variation in test scores than college attendance, and in their data, the school level ICCs for value added are .064 and .090 respectively.

[4] The MET data is geographically more urban than US schools as a whole, with a much higher percentage of African-American teachers; about 35% of MET teachers are black compared to

about 7% in the US (see e.g., NCES, 2017, Table 309.10). Our analytic sample of teachers includes: 16% Male, 56% White, 34% Black, 6% Hispanic and 2% Other. Teachers have 10.71 years of experience, and 7.7 years of district-specific experience, on average. 33% of teachers have a Master's degree or higher.

[5] In these data, the maximum likelihood estimated ICCs (the default in the mixed command) are nearly identical to estimates using a Bayesian approach, but are generally more conservative (lower) than ANOVA estimates of the ICC from STATAs 'loneway' command.

[6] We perform an exploratory factor analysis to determine the number of components that are actually separable in the data. CLASS has two factors with eigenvalues greater than 1, a possible rough rule of thumb for determining the number of factors. While FFT and PLATO have one factor that meets this criteria, both have multiple factors explaining a substantial portion of the variance, which is an additional criteria used to determine the number of factors. The first FFT factor explains 74% of the variance, while the second factor explains an additional 14%. The first PLATO factor explains 62% of the variance, while the second and third factors explain 20% and 17% of the variance, respectively. We perform an oblique rotation of the factors and take averages of the clearly separable groupings which emerge.

[7] Instruction as measured here is inherently a class/teacher-level phenomenon, rather than a student-level phenomenon, so from a measurement standpoint the ICCs in this analysis are not directly comparable to a student vs. school decomposition. In principle the Tripod is an exception, and derives from student-level reports, but the analysis here uses a class-level aggregation at level 1 (as reported by MET in the section-level analytic files).

[8] Ancillary analyses reveal that the greater relative variability in behavioral climate is not necessarily the case for the Tripod. However, students may have a more "myopic" view of behavioral climate than teachers or outside observers.

[9] The VAM estimates in MET, as well as alternative achievement measures, are provided only at the teacher level and thus are not comparable to prior research using student-level data.

[10] Clotfelter et al.'s analysis of 5th grade data from North Carolina (e.g., Tables 3 & 4) suggests it is possible for within-school teacher tracking to be minimal in a given sample of schools, about 45% of schools showed no evidence of within-school sorting on their set of measures.