education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 27 Number 7 January 28, 2019 ISSN 1068-2341

Complexity and Scale in Teaching Effectiveness Research: Reflections from the MET Study

Bryant Jensen
Brigham Young University

Tanner LeBaron Wallace
University of Pittsburgh

Matthew P. Steinberg
University of Pennsylvania

Rachael E. Gabriel
University of Connecticut

Leslie Dietiker
Boston University

Dennis S. Davis
North Carolina State University

Benjamin Kelcey
University of Cincinnati

Elizabeth Covay Minor
National Louis University

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 6/11/2018
Revisions received: 11/4/2018
Accepted: 11/4/2018

Peter Halpin
University of North Carolina at Chapel Hill
&
Ning Rui
Westat, Inc.
United States

Citation: Jensen, B., Wallace, T. L., Steinberg, M. P., Gabriel, R. E., Dietiker, L., Davis, D. S., … Rui, N. (2019). Complexity and scale in teaching effectiveness research: Reflections from the MET Study. *Education Policy Analysis Archives, 27*(7). http://dx.doi.org/10.14507/epaa.27.3923

Abstract: Researchers and policymakers in the US and beyond increasingly seek to identify teaching qualities that are associated with academic achievement gains for K-12 students through effectiveness studies. Yet teaching quality varies with academic content and social contexts, involves multiple participants, and requires a range of skills, knowledge, and dispositions. In this essay, we address the inescapable tension between complexity and scale in research on teaching effectiveness. We provide five recommendations to study designers and analysts to manage this tension to enhance effectiveness research, drawing on our recent experiences as the first external analysts of the Measures of Effective Teaching (MET) study. Our recommendations address conceptual framing, the measurement of teaching (e.g., observation protocols, student surveys), sampling, classroom videoing, and the use and interpretation of value-added models.
Keywords: teaching effectiveness; teaching quality; study design; secondary analysis

Complejidad y escala en la investigación de la eficacia de la enseñanza: Reflexiones del Estudio MET

Resumen: Investigadores y legisladores en los Estados Unidos y en otros países buscan identificar las cualidades de la enseñanza que se asocian con incrementos de desempeño académico para alumnos de primaria y secundaria a través de estudios de eficacia. Sin embargo, la calidad de la enseñanza varía según el contenido académico y los contextos sociales, involucra a múltiples participantes y requiere una variedad de habilidades, conocimientos y disposiciones. En este ensayo, abordamos la ineludible tensión entre la complejidad y la escala en la investigación sobre la eficacia de la enseñanza. Proveemos cinco recomendaciones a los diseñadores de estudios y analistas para manejar esta tensión y mejorar la investigación de eficacia, aprovechando nuestras experiencias recientes como los primeros analistas externos del estudio *Measures of Effective Teaching* (MET). Nuestras recomendaciones abordan el marco conceptual, la medición de la enseñanza (por ej., protocolos de observación, encuestas de estudiantes), el muestreo, el video en el aula y el uso e interpretación de modelos de valor agregado.
Palabras-clave: calidad de la enseñanza; eficacia; diseño de estudio; análisis secundario

Complexidade e escala na investigação da eficácia do ensino: Reflexões do estudo MET

Resumo: Pesquisadores e legisladores nos Estados Unidos e em outros países buscam identificar as qualidades de ensino associadas ao aumento do desempenho acadêmico de alunos do ensino fundamental e médio por meio de estudos de eficácia. No entanto, a qualidade do ensino varia de acordo com o conteúdo acadêmico e os contextos sociais,

envolve múltiplos participantes e requer uma variedade de habilidades, conhecimentos e disposições. Neste ensaio, abordamos a tensão inescapável entre complexidade e escala na pesquisa sobre a eficácia do ensino. Fornecemos cinco recomendações para projetistas e analistas de estudo para gerenciar essa tensão e melhorar a pesquisa sobre eficácia, alavancando nossas experiências recentes como os primeiros analistas externos do estudo *Measures of Effective Teaching* (MET). Nossas recomendações abordam a estrutura conceitual, a medição do ensino (por exemplo, protocolos de observação, pesquisas com estudantes), amostragem, vídeo em sala de aula e o uso e interpretação de modelos de valor agregado.
Palavras-chave: qualidade do ensino; eficácia; desenho do estudo; análise secundária

Introduction

Teachers and teaching are once again at the center of our national discourse about opportunity gaps in student achievement (e.g., Majoo, 2013). And while what occurs in classrooms is only one of many sets of factors that bear on student learning opportunities; among school variables, teaching quality accounts for the largest amount of variation in achievement (Rivkin, Hanushek & Kain, 2005). Hanushek (2016) explains that "a top teacher can in one year produce an added gain from students of one full year's worth of learning [...]" (p. 25).

Accounting for general sources in student achievement variation, however, is not synonymous with offering practical resolutions for improvement. Given the effects of teaching quality on student achievement, some argue for greater workforce management through summative evaluations of teachers to improve teaching (e.g., performance-based pay). This perspective, espoused by some "teacher evaluation" researchers, focuses on teacher behavior, while minimizing other classroom factors (e.g., curricula quality, student background, peer interactions) that also have been shown to bear on achievement (e.g., NRC, 2004). Others interpret the effects of teaching quality on student achievement by accounting for a wider array of factors, and pursue improvement through formative evaluations and practical frameworks that address differences in the content and the contexts of teaching (Gallimore & Santagata, 2006).

In this essay, we provide recommendations to enhance the latter body of research—"teach*ing* effectiveness" rather than "teach*er* effectiveness" or "teach*er* evaluation." Our purpose is to provide guidelines for designers of small-to-large-scale studies of teaching effectiveness, as well as to analysts who conduct effectiveness research using existing datasets. We base our recommendations on our recent, collective experience with the Measures of Teaching (MET) study, and illustrate these recommendations with specific examples from MET.

We define research on teaching effectiveness as generalizable studies of practical frameworks of teaching qualities that affect student achievement gains in particular subjects, across school and classroom contexts. Teacher effectiveness, on the other hand, examines how attributes of a teacher (e.g., credentials, background, beliefs) affect achievement gains. Researchers who examine the effectiveness of teachers in hiring, promotion, or other high-stakes decisions, as mentioned above, conduct "teacher evaluation" research.

We identify with Shulman's (1987) assertion that teaching is "outrageously complex" (p. 11). Day in and out, teachers are expected to plan lessons, organize instructional activities, foster warm and respectful relationships, sustain student interest, and partner with parents—in ways that promote learning for diverse students. Teaching involves not only what teachers do (e.g., Gage & Needles, 1989) and know (e.g., Ball, Thames & Phelps, 2008), but also who they and their students are in terms of dispositions (e.g., Borko, Liston & Whitcomb, 2007; Gay, 2015), identities (e.g., Akkerman & Meijer, 2011), and backgrounds (e.g., Zumwalt & Craig, 2005).

Providing generalizable evidence that reflects the complex nature of teaching is difficult. Effective teaching can vary by academic content (Grossman, Cohen, Ronfeldt & Brown, 2014) and the sociocultural contexts of classrooms (Tharp, 1989). The tension between (a) the complex nature of teaching and (b) scale in effectiveness research is inescapable. Scale—generalizability of findings due to statistical power—tends to reduce nuance in teaching, yet incorporating every covariate of interest into the design of effectiveness studies is not feasible. Large-scale studies, in other words, often do not capture the complex qualities of teaching that affect student learning, whereas small-scale studies tend to elucidate rich teaching constructs without providing generalizable evidence. Rather than dismiss this tension, we submit that there is value in exploiting it—that scale and complexity are dual imperatives in teaching effectiveness research. Failing to address this tension perpetuates our condition whereby policies about teaching are enacted with inadequate conceptualization or evidence (Darling-Hammond et al., 2010).

The MET Study

We draw on our recent experiences analyzing data from the Measures of Effective Teaching (MET) study to recommend five ways of managing tensions between complexity and scale. Our recommendations address the use of existing datasets for secondary analysis, as well as the design of new studies of teaching effectiveness, however large or small. The MET study was an ambitious, high-profile effort, funded by the Bill & Melinda Gates Foundation, to identify elements of teaching associated with academic gains of fourth- through ninth-grade students in math, biology, and English Language Arts. MET researchers followed 2,741 teachers in six school districts[1] over a two-year period, from 2009 to 2011. Multiple measures were administered on instructional practices, student achievement, as well as on the backgrounds and perceptions of teachers, students, and administrators. Student achievement gains formed the basis for calculating value-added models assigned to teachers, and panoramic videos of classroom lessons provided source material to observe and score instructional practices using a series of protocols.

The stated purpose of the MET study, commissioned rather than resulting from a blind-peer-review process, was to "test new approaches to measuring teaching" (Kane & Cantrell, 2010, p. 2), not necessarily to provide new insights about effectiveness. The study was more concerned with predictive than construct validity (Martínez, Schweig & Goldschmidt, 2016). Thus, some validity and reliability standards were prioritized over others in the selection of measures. We were the first external users to analyze MET data. Through a competition by the National Academy of Education, we were awarded ten early career grants to pursue original studies using data stored in the MET Longitudinal Database (MET LDB). Our projects were diverse (e.g., micro-analyses of instruction, estimation of teaching effects on student performance, developing new measures of teaching), and provided us with varied perspectives on the study and its data.

There are a number of large-scale studies of teaching effectiveness that could have been used as a basis for this essay, such as the 1999 Third International Mathematics and Science Study (TIMSS; Stigler & Hiebert, 2009). We chose MET because of (a) our recent experiences with these data, (b) its unprecedented size in terms of sampled classrooms and the variety of data collected, and (c) its explicit focus on "effective teaching." To provide the recommendations that follow, we

[1] Districts included Charlotte-Mecklenburg Schools (NC), Dallas Independent School District (TX), Denver Public Schools (CO), Hillsborough County Public Schools (FL), Memphis City Schools (TN), and the New York City Department of Education (NY).

examine the relative strengths and weakness of the MET study design. We do not intend to encourage or to discourage use of MET data by analysts or others, but to use MET as a case to illustrate our recommendations to enhance future research on teaching effectiveness.

Our Recommendations

In our deliberations about these recommendations, we aimed to strike a balance—and to identify relationships—between technical, empirical, and conceptual issues in teaching effectiveness research. We seek to build on recommendations provided elsewhere (see, for example, special issues on value-added models in *Education Policy Analysis Archives* [Amrein-Beardsley, Collins, Polasky & Sloat, 2013] and *Educational Researcher* [Harris & Herrington, 2015]) by attending to the complexity-scale tension identified, in the context of the MET study.

Recommendation 1: Designs of large-scale studies should reflect a clear and coherent conceptualization of teaching effectiveness

Any study of teaching should provide an explicit framework with associated claims. In effectiveness research, this includes specifying the unit(s) of analysis—e.g., teacher actions, teacher knowledge, curricula, teacher-student interactions, peer interactions—as well as providing theory-based conjectures about how and why specific aspects of teaching affect student learning or development. Analytic units should be congruent with—and conjectures verifiable from—the measures of teaching incorporated in a given study.

The MET study used multiple types of teaching measures, including observation protocols, value-added models, teacher surveys, and student surveys.[2] Some addressed generic dimensions of teaching quality, whereas others captured content-specific dimensions (Kane & Staiger, 2012). It is not clear how these various measures and corresponding constructs, however, were intended to operationalize an underlying framework of teaching quality. This could reflect the econometric approach to effectiveness research upon which the MET study was based, which emphasizes inputs and outputs while attending less to matters of process. Indeed, the stated purpose of MET was to assess the validity and reliability of various measures of teaching in a large sample, to describe effectiveness in terms of these measures, and to determine whether multiple scores across measures can "be combined to develop a set of fair, valid, and reliable indicators of teaching quality for use in teacher evaluation systems intended to rank teachers for personnel decision making" (White & Rowan, 2014, p. 5). It was designed as an empirical/policy exercise, not necessarily framed as a conceptual exercise of teaching effectiveness. The unit of analysis simply was the teacher, rather than detailing conjectures about how specific qualities of teaching affect student learning in a particular content or social context.

As an example, MET designers did not specify how teachers' "content knowledge for teaching" (CKT; Ball, Thames & Phelps, 2008) would (a) underlie their instructional practice or (b)

[2] Observation protocols included the Classroom Assessment Scoring System (CLASS), the Framework for Teaching (FFT), and abbreviated versions of the Mathematical Quality of Instruction (MQI), the Protocol for Language Arts Teaching Observations (PLATO), and the Quality of Science Teaching (QST). The CLASS and FFT are generic measures of instructional quality, whereas the MQI, PLATO, and QST are content-specific measures (mathematics, English/Language Arts, and science, respectively). The UTeach Observation Protocol (UTOP) was also used to score a small subset of math and science classroom videos—teachers certified by the National Board of Professional Teaching Standards. See Tables A2 through A4 in the MET Study User's Guide (White & Rowan, 2014, pp. 60-64) for the number of teachers and classrooms per protocol that were assigned to be scored for years one and two.

affect how well students learned a specific content. Yet several CKT items were included in the teacher survey, along with items about the quality of working conditions and perspectives on local teacher evaluation practices. After CKT items were found to be statistically uncorrelated with student achievement gains, "the MET project did not include CKT results within its composite measure of effective teaching" (Cantrell & Kane, 2013, p. 15). It is unclear which conjectures were tested in this analysis, or how theories of effective teaching might be reconsidered or revised based upon the lack of significant relationships.

In addition, in the design of MET, several classroom observation protocols were used without an underlying rationale. The protocols, for example, emphasize teacher over student behavior, even when peer supports constitute strong indicators of teaching effectiveness (Wallace, Sung & Williams, 2014; Webb et al., 2008). Similarly, the protocols are limited in their ability to address contextual or cultural dimensions of teaching (Jensen, Grajeda & Haertel, 2018), or the ways teachers differentiate instruction from one student to another (e.g., Connor et al., 2011). All studies must make tradeoffs and choose areas of focus, so decisions about measures used in MET are not inherently problematic; the concern is a lack of conceptual frame to justify these decisions.

Whereas using multiple measures of teaching in effectiveness research can provide analytic advantages (Martínez, Schweig & Goldschmidt, 2016)—including the use of latent class analysis to provide diagnostic information to teachers to improve their practice (Halpin & Kieffer, 2015)—a focus on scale should not disregard matters of complexity. Though "urban" was used to characterize participating schools in the study, for example, it is unclear how MET measures account for well-documented considerations specific to urban education like racial segregation, community partnerships, or student resistance (e.g., Milner, 2012). From the outset, study designs should be intentional and explicit about framing aspects of teaching and learning, provide specific conjectures, and align measures of teaching within this framework. Balancing scale with complexity in this way allows effectiveness research to address how and why qualities of teaching affect achievement domains across different contexts, not simply to rank teachers or to make wholesale statements about "what works."

Highlighting a lack of conceptual framing in the MET study is not necessarily a statement about the usefulness of MET data to analysts. It means that (a) some teaching and learning constructs across MET measures are theoretically incongruent, and (b) analysts have to spend substantial time in the user's guide (White & Rowan, 2014) and with MET files to know whether the data can be used to address their research questions. Davis, Bippert and Villarreal (2015), for example, spent several months with MET data before concluding that the full range of teaching practices to support students' reading comprehension was not present in the English Language Arts lessons. We assert that there is no such thing as neutrality or "conceptual agnosticism" in teaching effectiveness research. The best way to build a useful repository of teaching effectiveness data is to provide a clear and coherent framework from the outset.

Recommendation 2: Measures of teaching used in large-scale studies should be technically strong

No single survey, observation protocol, performance assessment, or artifact-based measure of teaching can capture its complexity, and "tacit judgments and dilemmas [are] embedded in [all] measures [of] teaching" (Darling-Hammond et al., 2010, p. 88). Combining constructs across measures should be done in ways that enhance validity and maximize score reliability rather than "seek to optimally predict student test scores" (Martínez, Schweig & Goldschmidt, 2016, p. 738). Once selected for clear and coherent reasons, the technical properties of teaching measures should be scrutinized in at least two stages.

First, there should be evidence that measures meet basic validity and reliability standards before incorporating them into large-scale studies. Though used widely, some measures in the MET study did not meet these standards before selecting them. Adapted from Ferguson's (2012) measure, for example, the Tripod is a student survey designed to gather student perceptions of teaching quality along seven dimensions. Analyses of Tripod data in the MET showed that student perceptions of teacher strengths and weaknesses were (a) fairly consistent across the classrooms they taught and (b) significantly associated with student achievement gains (Kane & Cantrell, 2010). Yet the internal structure (i.e., construct validity) of the Tripod was not established before the MET study, which muddled interpretations and implications of MET findings. Rather than a seven-factor structure, Wallace, Kelcey and Ruzek (2016) found a bi-factor structure consisting of a general instructional quality factor and classroom management-specific factor to best fit MET Tripod data from mathematics classrooms in grades 6 through 8. Additional work is needed to examine relationships between student perception and other measures of teaching quality.

Second, developers should pilot the technical properties of teaching measures in ways that approximate the conditions of—and prior to—the large-scale study design. Though all but one of the classroom observation protocols used in the MET study were vetted in blind, peer-reviewed outlets before the MET study; by and large, they were not tested in ways that approximated its large-scale design, in which video segments were assigned at random to a large number of trained raters (*n* = 902) to score with generic and content-specific protocols.[3] This, among other possibilities, likely contributed to low reliabilities of observation scores. The amount of score variation explained by differences among teachers was lower than expected—14 to 37 percent across instruments, which meant that best-case-scenario reliability coefficients ranged from .39 to .67 (Kane & Staiger, 2012). This suggests that MET observation scores were "largely driven by factors other than consistent aspects of a teacher's practice" (p. 17).

Addressing the technical properties of teaching measures before incorporating them into effectiveness studies enhances the usefulness of study findings. It affords opportunities to revise the design, to alter the measures themselves, or both. There should be a process of adapting—rather than simply adopting—measures of teaching to large-scale studies. MET researchers, for example, could have examined how assigning classroom videos to many raters, with a single rater per video—rather than double-coded videos by a handful of raters—affected the reliability of observation scores. Analyzing rater thinking while scoring could have informed rubric revisions to reduce inference demands and, thus, increase reliabilities (Bell et al., 2014; Kelcey, McGinn & Hill, 2014).

Recommendation 3: Minimize sampling problems, but make the most of those that happen to occur

Sampling is the single-most challenging design feature of large-scale effectiveness studies. In addition to promoting external and internal validity through random selection and assignment, decisions about how to sample teachers and students should anticipate and make the most of practical problems like attrition and non-compliance with random assignment protocols. Doing so can provide new insights about teaching effectiveness.

The MET sample provides opportunities to study timely issues of excellence and equity in teaching. A large portion of MET teachers identified as persons of color (44%, compared to 17% nationally; see Table 1), and the diversity of student composition in classrooms from six large, urban

[3] Rater calibration was ongoing, and "scoring leaders" provided regular assistance (White & Rowan, 2014, p. 30). Some design features of the MET study were piloted in the Understanding Teaching Quality (UTQ) study (see http://utqstudy.org/index.html).

districts affords analyses about equity issues (e.g., differential effects of instructional quality; Covay Minor, 2015) that are not possible in other large-scale studies.

Yet, MET teachers were a convenience sample. They volunteered to participate with compensation[4], which constrains inferences from MET analyses. Generalizability is further inhibited by teacher attrition from Years 1 to 2 (see Table 1).[5] Overall attrition was 31 percent, and rates varied by district. Stated reasons for attrition included schools and teachers opting out, teacher mobility, teachers moving grades or subjects, and teacher illness (White & Rowan, 2014, p. 16). Though MET teachers did not differ significantly from non-MET teachers in terms of race, ethnicity, and years of teaching experience (Kane & Staiger, 2012); it is not clear which of the above reasons for attrition were the most dominant, or how exactly they biased the sample.

Additional attrition is found in classroom videos for secondary analysis. More than a third of MET teachers did not re-consent their videos to be used for secondary analysis, and additional videos that were consented are not currently available. At the time of writing this essay, 21.7 percent of all videos are available for secondary analysis.

The effort to randomly assign teachers to classrooms in the second year of the MET study is also noteworthy. Controlling for the bias of non-random teacher assignment affords analyses of causal questions regarding teaching effectiveness. For example, the randomization of teachers to classes enabled an assessment of the extent to which individual (Garrett & Steinberg, 2015) and composite (Kane, McCaffrey, Miller, & Staiger, 2013) measures of teaching quality could identify "effective" teachers. Teachers, not individual students, were assigned randomly to classes. MET data enable estimates of teaching effectiveness for groups of students, but not to compare teachers for any given student.

Rates of student compliance in the randomization process ranged from 27.4 to 65.6% across districts (see Table 2 in Kane et al., 2013).[6] Non-compliance creates challenges for identifying causal effects of teaching, though it can also create opportunities for other analyses. Steinberg and Garrett (2016), for example, show that MET classroom observation scores were biased by the nonrandom sorting of classes to teachers, which occurred after the random assignment of teachers. This implied that conditioning teaching scores on students' incoming achievement was insufficient to compare or to make inferences about teachers. They concluded that caution should be taken when making high-stakes personnel decisions based on classroom observation scores. This analysis and associated conclusions would not have been possible with full compliance of the randomization sample. Thus, analysts should be creative in taking advantage of unexpected sampling blunders, inevitable to some degree in large studies of teaching.

[4] Teachers in sampled schools were offered $1,500 to participate in the study ($1,000 at the beginning and $500 at the end of the study), in addition to small gifts from district-MET budgets.

[5] Page 23 of the MET Study User's Guide (White & Rowan, 2014) states that the full Y2 teacher sample was 2,086. The core Y2 sample (those with data both years of the study) in Table 3 on page 24 adds to 1,902. The latter number is reiterated in Appendix A (page 58). This discrepancy is due to the number of teachers who participated during Y2 only (*n*=184).

[6] A total of 865 exchange groups in 316 schools were requested, whereas 619 groups in 284 schools were actually assigned (Kane, McCaffrey, Miller & Staiger, 2013, p. 14)

Table 1
Teacher Demographics in Years 1 and 2 of the MET Study

		Year 1 (*N* = 2,741)	Year 2 (*N* = 1,902)
Gender	Female	2,077 (75.8%)	1,445 (76.0%)
	Male	644 (24.2%)	457 (24%)
Race/Ethnicity	White	1,537 (56.1%)	1,073 (56.4%)
	Black	829 (30%)	569 (30%)
	Hispanic	154 (5.6%)	106 (5.5%)
	Other	78 (2.8%)	56 (3.0%)
Grade	4	436 (15.9%)	277 (14.5%)
	5	438 (16.0%)	310 (16.3%)
	6	443 (16.1%)	307 (16.1%)
	7	376 (13.7%)	264 (13.9%)
	8	327 (11.9%)	269 (14.1%)
	9	715 (26.1%)	479 (25.2 %)
Subject	Multiple	5 (.1%)	1 (< .1%)
	ELA	980 (35.7%)	722 (40.0%)
	Math	863 (31.4%)	652 (34.2%)
	ELA + Math	658 (24.0%)	365 (19.2%)
	Biology	240 (8.8%)	163 (8.6%)
Years taught		Median = 7, IQR = (3.0, 9.7)	Median = 7, IQR = (3.0, 9.9)
Year taught in current district		Median = 5 IQR = (2.7, 7.3)	Median = 5, IQR = (2.8, 7.2)
Master’s degree or higher		728 (26.5%)	498 (26.1%)

Note: Total sample sizes for each year were for teachers had at least one class section active in the MET study. IQR denotes interquartile range.

Whereas sampling problems like attrition and non-compliance are not altogether avoidable, and MET researchers should be commended for designing a study that allows for causal inferences, a few suggestions could help minimize sampling problems in future teaching effectiveness studies, large or small. These recommendations can be implemented with minimal cost or technology demands.

First, researchers should collaborate with teachers to conceptualize the purpose of study. Researchers can draw on teachers' "wisdom of practice" (Shulman, 2004) to refine core conjectures of teaching effectiveness upon which to build the study design. Explaining the purpose of study to teachers, in the language of teachers, may help with recruitment and to reduce attrition problems. Second, and relatedly, researchers can provide teachers with formative feedback after study completion, which can also help with recruitment and retention. Lastly, gather teacher consent from the outset for external research use of videos. This can reduce video selection bias for external researchers who wish to use videos for various analytic purposes.

Recommendation 4: Use videoing procedures that provide as much insight as possible into conjectures about teaching effectiveness

Video data have a long and rich history in research on teaching (Erickson, 2011). They provide analysts with multiple opportunities to examine teacher and student actions and interactions from a variety of perspectives to draw interpretations and conclusions. Yet decisions about how and what to video record have implications for the types of analyses that can be conducted. Several decisions about classroom videoing (e.g., when to record, the type and number of cameras, the length and number of segments, how to capture audio) can afford or constrain analytic possibilities (Derry et al., 2010). To maximize their utility, videos should capture the specified unit(s) of analysis, as well as teaching activities that are most relevant to researchers' core conjectures about effectiveness.

The MET study included a herculean video effort. At least four videotaped lessons (two scored segments per lesson) were recorded for each teacher in the longitudinal sample, using 360-degree panoramic cameras that captured most students in the classroom. Videos include lessons in mathematics, English language arts, and biology (see Table 1), and content markers are included in video IDs to facilitate identification (White & Rowan, 2014, pp. 46-48). Panoramic views were often paired with a video recordings of the front of the classroom, typically capturing a white board or a smart board. Combining video captures with a range of other data sources, including images of classroom artifacts (e.g., worksheets), affords an exciting range of analytic possibilities. Linking videos to data on teaching quality, student development, and contextual variables allows for a variety of analyses not possible in other studies. Richman, Dietiker, & Riling (2018), for example, analyzed a lesson with high student interest in mathematics to explain how the unfolding mathematical ideas supported the reactions of excitement demonstrated by students in the video.

Decisions about how and what to record in MET videos also constrain analytic possibilities. We identify two constraints. First, the teachers chose which lessons to record. Though MET teachers were asked by researchers to record two lessons of "focal topics" and another two while teaching a topic of choice, they decided for themselves when and what to record. As is often the case when being evaluated, it could be that MET teachers prioritized whole-group, teacher-directed lessons of content perceived to be more "academically rigorous," foregrounding "on-stage" teaching (Sawyer, 2004). In their analysis of reading comprehension instruction in grades 4-8, Davis, Bippert and Villarreal (2015) found that in most MET lessons (i.e., 96% of ELA lessons that emphasized text comprehension), teachers worked in whole-class arrangements—presenting lengthy content, facilitating whole-group activities, and moving around the room to manage student work. Yet this format is one of many instructional arrangements teachers use for reading comprehension instruction. The high prevalence of whole-group, on-stage teaching could be an artifact of the data collection process.

Second, stationary cameras and audio-recording in MET videos constrain analytic possibilities. Using stationary cameras rather than videographers or cameras with automated zoom features meant that most students in most MET videos are too far away to decipher what they are doing or saying, especially during group work. Only two microphones, one worn by the teacher and the other fixed to the camera cart, were used. MET video files include an audio quality variable (high, mid, low) at the segment level, which is helpful for constructing analytic samples, but most student talk in videos with "high quality" audio is still difficult to decipher. Moreover, filtering segments by audio quality imposes another layer of possible selection bias.

Given limitations imposed by the ways classroom lessons were selected and captured (video *and* audio), MET videos are best used for examining research questions regarding teacher talk and behavior rather than what the students do or say. Designs of future teaching effectiveness studies can manage the scale-complexity tension to address these limitations by (a) randomly selecting days and times to record lessons, (b) aligning videoing procedures with specific frameworks for teaching effectiveness and associated claims (see Recommendation 1), and (c) enhancing student audio capture with additional accessory microphones.

Recommendation 5: Be cautious and nuanced in using and interpreting VAM scores

Value-added models (VAM) are used to isolate the effects of classrooms on student achievement by analyzing test score gains and applying statistical controls. A series of assumptions and limitations apply to all VAM scores (AERA, 2015; Kane, 2017; Haertel, 2013). To manage the complexity-scale tension in teaching effectiveness research, conclusions based on VAM analyses should reflect the specifics of each VAM construction, and acknowledge how VAM-based rankings "var[y] across models, courses, and years" (Darling-Hammond et al., 2010, p. 91).

White and Rowan (2014, pp. 27-28) detail VAM procedures in the MET study—how the construction of VAM scores varied by district, test (state measures versus MET supplements), and grade. Student achievement in all cases was converted to rank-based z-scores for VAM computation, though some tests were ranked within districts and others were ranked across districts. Moreover, student-level covariates (e.g., ethnicity, ELL status, free/reduced lunch, etc.) used to estimate VAM scores varied by district, and VAM scores using supplemental tests did not control for incoming student performance, as they did with state exams (White & Rowan, 2014, p. 28). These inconsistencies explain why VAM correlations between state and supplemental tests vary widely by district and academic content in MET. When using VAM scores to compare classroom effects across sites, users should make sure data are consistent with corresponding assumptions (Baker et al., 2010).

Sampling bias gives reason for further caution to analysts who wish to use and interpret MET VAM scores. Many of the MET students (i.e., a fourth to a fifth across classrooms) who were supposed to take supplemental tests did not, which biased the estimates. In addition, VAM scores in MET do not account for school fixed effects. They control for student background, but not for between-school mean differences. This further biases VAM estimates because student populations served by schools vary greatly within districts (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). The MET database has enough information to construct VAM score adjustments that meet external researchers' needs, which is what we advise. We also recommend using multiple measures in conjunction with VAM scores to triangulate findings (Darling-Hammond, 2015; Johnson, 2015).

VAM scores are compelling, especially for policy, because they aim to isolate classroom effects on student achievement. Analyses of VAM scores in the MET study indicate that teachers' past record of value-added is a strong predictor of student achievement across content areas and school years, and that classrooms have larger effects on math than reading performance (Kane & Cantrell, 2010). Yet, researchers should consider the construction of specific models before using VAMs in their secondary analyses.. Streamlining VAM construction and preventing sampling bias in large studies is difficult to impossible. The larger the study, the more difficult it is to control the quality of VAM scores—yet another example of the tension between scale and complexity in teaching effectiveness research.

Conclusion

A primary purpose of studying relationships between qualities of teaching and student achievement gains is to improve learning opportunities in classrooms. Improvement requires researchers, policymakers, and educators (a) to account for the complexities inherent in teaching and (b) to provide generalizable evidence at scale. We have shared five recommendations to enhance teaching effectiveness research by managing—rather than disregarding—the tensions that arise between these two demands. Illustrations of our recommendations are also useful to researchers who consider using MET data for secondary analyses.

Designs of effectiveness studies—and the measures and methods that they use—should be conceptually clear and coherent, technically strong, and transparent about practical problems that invariably arise in the data collection process. This way, effectiveness studies can identify the operative mechanisms affecting how students learn specific academic content across a variety of classroom contexts. We can learn a great deal from past efforts, like the MET study, about how to test specific conjectures regarding teaching effectiveness; to sample students, teachers, and schools strategically; to develop measures thoughtfully; and to interpret our findings carefully. This way education policies can follow nuanced evidence to scale improvements.

Acknowledgements

We thank reviewers for their useful feedback to improve this manuscript. We also thank our colleagues and mentors for lengthy discussions that led to this essay, as well as readers who gave us insightful feedback before submission. In particular, we want to acknowledge and express our gratitude to Robert V. Bullough, Edward Haertel, José Felipe Martínez, and Brian Rowan.

References

Akkerman, S. F., & Meijer, P. C. (2011). A dialogical approach to conceptualizing teacher identity. *Teachers and Teacher Education, 27*(2), 308-319. https://doi.org/10.1016/j.tate.2010.08.013

American Educational Research Association (2015). AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452. https://doi.org/10.3102/0013189X15618385

Amrein-Beardsley, A., Collins, C., Polasky, S. A., & Sloat, E. F. (2013). Value-added model (VAM) research for education policy: Framing the issue. *Education Policy Analysis Archives, 21*(4). http://dx.doi.org/10.14507/epaa.v21n4.2013

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R., …Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper No. 278). Washington, DC: Economic Policy Institute.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching what makes it special? *Journal of Teacher Education*, *59*(5), 389-407. https://doi.org/10.1177/0022487108324554

Bell, C. A., Qi, Y., Croft A. J., Leusner D., McCaffrey D. F., Gitomer, D. H., & Pianta, R. C. (2014). Improving observational score quality: Challenges in observer thinking. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project*. San Francisco, CA: Jossey-Bass.

Borko, H., Liston, D., & Whitcomb, J. A. (2007). Apples and fishes: The debate over dispositions in teacher education. *Journal of Teacher Education*, *58*(5), 359-364. https://doi.org/10.1177/0022487107309977

Cantrell, S., & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Seattle, WA: Bill & Melinda Gates Foundation.

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P., … Schatschneider, C. (2011). Classroom instruction, child X instruction interactions and the impact of differentiating student instruction on third graders' reading comprehension. *Reading Research Quarterly*, *46*(3), 189-221.

Covay Minor, E. (2015). Classroom composition and racial differences in opportunities to learn. *Journal of Education for Students Placed at Risk, 20*(3), 238-262. https://doi.org/10.1080/10824669.2015.1043009

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? *Educational Researcher, 44*(2), 132-137. https://doi.org/10.3102/0013189X15575346

Darling-Hammond, L., Dieckmann, J., Haertel, E., Lotan, R., Newton, X., Philipose, S., … Williamson, P. (2010). Studying teacher effectiveness: The challenges of developing valid measures. In G. Walford, E. Tucker & M. Viswanathan (Eds.), *The SAGE handbook of measurement* (pp. 87-106). Thousand Oaks, CA: SAGE Publishing. https://doi.org/10.4135/9781446268230.n6

Davis, D. S., Bippert, K., & Villarreal, L. (2015). Instructional tendencies in the teaching of reading comprehension: A portrait of practice in the Measures of Effective Teaching (MET) database. *Literacy Research: Theory, Method, and Practice, 64*, 285-306. https://doi.org/10.1177/2381336915617399

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., ... & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences*, *19*(1), 3-53. https://doi.org/10.1080/10508400903452884

Erickson, F. (2011). Uses of video in social research: a brief history. *International Journal of Social Research Methodology*, *14*(3), 179-189. https://doi.org/10.1080/13645579.2011.563615

Ferguson, R. (2012). Can student surveys measure teaching quality? *Kappan, 94*(3), 24-28. https://doi.org/10.1177/003172171209400306

Gage, N. L., & Needels, M. C. (1989). Process-product research on teaching: A review of criticisms. *The Elementary School Journal, 89*(3), 253-300. https://doi.org/10.1086/461577

Gallimore, R., & Santagata, R. (2006). Researching teaching: The problem of studying a system resistant to change. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research* *methodology: Psychological measurement and evaluation* (pp. 11-28). Washington, DC: American Psychological Association. https://doi.org/10.1037/11384-001

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores. *Educational Evaluation and Policy Analysis, 37*(2), 224-242. https://doi.org/10.3102/0162373714537551

Gay, G. (2015). Teachers' beliefs about cultural diversity. In H. Fives & M. Gill (Eds.), *International handbook of research on teachers' beliefs*, 453-474. New York: Routledge.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293-303.

Haertel, E. (2013). *Reliability and validity of inferences about teachers based on student test scores.* Princeton, NJ: Educational Testing Service.

Halpin, P. F., & Kieffer, M. J. (2015). Describing profiles of instructional practice: A new approach to analyzing classroom observation data. *Educational Researcher*, *44*(5), 263-277. https://doi.org/10.3102/0013189X15590804

Harris, D. N., & Herrington, C. D. (2015). The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher, 44*(2), 71-76. https://doi.org/10.3102/0013189X15576142

Hanushek, E. A. (2016). What matters for student achievement. *Education Next*, *16*(2), 19-26.

Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment, 23*(4), 250-276. https://doi.org/10.1080/10627197.2018.1515010

Johnson, S. M. (2015). Will VAMs reinforce the walls of the egg-crate school? *Educational Researcher, 44*(2), 117-126. https://doi.org/10.3102/0013189X15573351

Kane, M. T. (2017). *Measurement error and bias in value-added models*. Princeton, NJ: ETS.

Kane, T., & Cantrell, S. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T., McCaffrey, D., Miller, T., & Staiger, D. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment. Research paper.* Seattle, WA: Bill & Melinda Gates Foundation.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teachers: Combining high quality observations with student surveys and achievement gains.* Seattle, WA: Bill & Melinda Gates Foundation.

Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate Measurement Invariance in Cross classified Rater-mediated Assessments. *Frontiers in Psychology*, *5*, 1-13. https://doi.org/10.3389/fpsyg.2014.01469

Majoo, F. (2014, September 3). Grading teachers, with data from class. *The New York Times.*

Martínez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*(4), 738-756. https://doi.org/10.3102/0162373716666166

McCaffrey, D., Lockwood, J. R., Koretz, D., Louis, T., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101. https://doi.org/10.3102/10769986029001067

Milner, H. R. (2012). But what is urban education? *Urban Education*, *47*(3), 556-561. https://doi.org/10.1177/0042085912447516

National Research Council. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, D.C.: National Academies Press.

Richman, A. S., Dietiker, L., & Riling, M. (online, 2018). The plot thickens: The aesthetic dimensions of a captivating mathematics lesson. *The Journal of Mathematical Behavior*. https://doi.org/10.1016/j.jmathb.2018.08.005

Rivkin, S., Hanushek, E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Sawyer, R. K. (2004). Creative teaching: Collaborative discussion as disciplined improvisation. *Educational Researcher*, *33*(2), 12-20. https://doi.org/10.3102/0013189X033002012

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Shulman, L. S. (2004). *The wisdom of practice: Essays on teaching, learning, and learning to teach.* San Francisco, CA: Jossey-Bass.

Steinberg, M., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis, 38*(2), 293- 317. https://doi.org/10.3102/0162373715616249

Stigler, J. W., & Hiebert, J. (2009). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom.* New York: Free Press.

Tharp, R. G. (1989). Psychocultural variables and constants: Effects on teaching and learning in schools. *American Psychologist, 44*(2), 349-359. https://doi.org/10.1037/0003-066X.44.2.349

Wallace, T. L., Sung, H. C., & Williams, J. D. (2014). The defining features of teacher talk within autonomy-supportive classroom management. *Teaching and Teacher Education*, (42), 34-46. https://doi.org/10.1016/j.tate.2014.04.005

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the Tripod student perception survey. *American Educational Research Journal, 53*(6), 1834-1868. https://doi.org/10.3102/0002831216671864

Webb, N. M., Franke, M. L., Ing, M., Chan, A., De, T., Freund, D., & Battey, D. (2008). The role of teacher instructional practices in student collaboration. *Contemporary Educational Psychology, 33,* 360-381. https://doi.org/10.1016/j.cedpsych.2008.05.003

White, M., & Rowan, B. (2014). *User guide to the Measures of Effective Teaching Longitudinal Database* (MET LDB). Ann Arbor, MI: Inter-University Consortium for Political and Social Research, The University of Michigan.

Zumwalt, K. & Craig, E. (2005). Teachers' characteristics: Research on demographic profile. In M. Cochran-Smith & Zeichner, K. M. (Eds), *Studying Teacher Education: The Report of the AERA Panel on Research and Teaching Education*. Mahwah, NJ: Lawrence Erlbaum.

About the Authors

Bryant Jensen
Brigham Young University
bryant_jensen@byu.edu

Bryant Jensen is an associate professor of teacher education at Brigham Young University. His research addresses equity in teaching and learning for children from minoritized communities, especially Latino children from immigrant families.

Tanner LeBaron Wallace
University of Pittsburgh
twallace@pitt.edu

Tanner LeBaron Wallace is an associate professor of applied developmental psychology at the University of Pittsburgh. Her research addresses race consciousness among white teachers, and how interpersonal connections form social contexts for motivation.

Matthew P. Steinberg
University of Pennsylvania
steima@upenn.edu

Matthew P. Steinberg is an assistant professor of education policy at the University of Pennsylvania. His research focuses on teacher evaluation and human capital, school discipline and safety, urban school reform, and school finance.

Rachael E. Gabriel
University of Connecticut
rachael.gabriel@uconn.edu

Rachael Gabriel is an associate professor of literacy education at the University of Connecticut. Her research is focused on literacy instruction, supports for adolescent literacy, state literacy policies, and teacher evaluation systems.

Leslie Dietiker
Boston University
dietiker@bu.edu

Leslie Dietiker is an assistant professor of mathematics education at the Wheelock College of Education & Human Development at Boston University. Her research addresses how the aesthetic dimensions of mathematics curriculum can impact student mathematical experiences.

Dennis S. Davis
North Carolina State University
ddavis6@ncsu.edu

Dennis Davis is an associate professor of literacy education at NC State University. His research focuses on reading comprehension and practices for supporting readers who have difficulties with literacy in school.

Benjamin Kelcey
University of Cincinnati
kelceybn@ucmail.uc.edu
Benjamin Kelcey is an associate professor of quantitative reasoning at the University of Cincinnati. His research focuses on causal inference and measurement methods within the context of multilevel and multidimensional settings such as classrooms and schools.

Elizabeth Covay Minor
National Louis University
eminor1@nl.edu
Elizabeth Covay Minor is an assistant professor of educational leadership at National Louis University. Her research focuses on inequality in student opportunities to learn.

Peter Halpin
University of North Carolina at Chapel Hill
peter.halpin@unc.edu
Peter Halpin is an associate professor of quantitative methods in the School of Education at the University of North Carolina at Chapel Hill. His research focuses on psychometrics (e.g., confirmatory factor analysis, item response theory, latent class analysis) and technology-enhanced assessments in education.

Ning Rui
Westat, Inc.
ningrui@westat.com
Ning Rui is a researcher at Westat, Inc. His research addresses quantitative methods, educational reform, and, in particular, value-added models.

education policy analysis archives

Volume 27 Number 7 January 28, 2019 ISSN 1068-2341