SPECIAL ISSUE
Policies and Practices of Promise in Teacher Evaluation

education policy analysis archives

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

epaa aape

Arizona State University

Volume 28 Number 62 April 13, 2020 ISSN 1068-2341

Using Global Observation Protocols to Inform Research on Teaching Effectiveness and School Improvement: Strengths and Emerging Limitations

Sean Kelly
University of Pittsburgh
United States

Robert Bringe
University of North Carolina-Chapel Hill
United States

Esteban Aucejo
Arizona State University
United States
&
Jane Fruehwirth
University of North Carolina-Chapel Hill
United States

Citation: 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).

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

Manuscript received: 8/31/2019
Revisions received: 1/15/2020
Accepted: 2/20/2020

https://doi.org/10.14507/epaa.28.5012 This article is part of the special issue, *Policies and Practices of Promise in Teacher Evaluation*, guest edited by Audrey Amrein-Beardsley.

Abstract: An essential feature of many modern teacher observation protocols is their "global" approach to measuring instruction. Global protocols provide a summary evaluation of multiple domains of instruction from observers' overall review of classroom processes. Although these protocols have demonstrated strengths, including their comprehensiveness and advanced state of development, in this analysis we argue that global protocols also have inherent limitations affecting both research use and applied school improvement efforts. Analyzing the Measures of Effective Teaching study data, we interrogate a set of five potential limitations of global protocols. We conclude by discussing fine-grained measures of instruction, including tools that rely on automated methods of observation, as an alternative with the potential to overcome many of the fundamental limitations of global protocols.
Keywords: Teacher Improvement; Instructional Methods; Observation; Educational Technology

Uso de protocolos de observación global en investigaciones sobre efectividad de los docentes y mejoramiento de escuelas: Fortalezas y limitaciones emergentes
Resumen: Una característica esencial de varios protocolos de observación es su aproximación "global" a la medición de métodos de instrucción. Dichos protocolos proveen una evaluación sumaria de múltiples aspectos de instrucción docente que son creados en base al análisis de observadores de clases. A pesar de que estos protocolos presentan fortalezas, incluyendo su compresivo análisis y avanzado estado de desarrollo, en este estudio argumentamos que dichos protocolos tienen limitaciones inherentes que afectan tanto a su uso para investigación como para la implementación de esfuerzos relacionados con el mejoramiento de escuelas. En base al análisis de datos generados por el estudio Medidas de Enseñanza Efectiva, exploramos cinco potenciales limitaciones de los protocolos globales. Concluimos el análisis discutiendo medidas detalladas de instrucción, incluyendo herramientas que se relacionan con métodos automatizados de observación, como una alternativa con el potencial para superar varias de las limitaciones fundamentales de los protocolos globales.
Palabras-clave: Mejoramiento Docente, Métodos de Instrucción, Observación, Tecnología Educativa

Uso de protocolos globais de observação em pesquisas sobre a eficácia do professor e melhoria da escola: pontos fortes e limitações emergentes
Resumo: Uma característica essencial de vários protocolos de observação é sua abordagem "global" para a medição de métodos instrucionais. Esses protocolos fornecem uma avaliação resumida de vários aspectos da instrução de ensino criados com base na análise dos observadores de classe. Embora esses protocolos tenham pontos fortes, incluindo sua análise compressiva e status de desenvolvimento avançado, neste estudo, argumentamos que esses protocolos têm limitações inerentes que afetam tanto seu uso em pesquisas quanto na implementação de esforços relacionados à melhoria da escola. Com base na análise dos dados gerados pelo estudo Effective Teaching Measures, exploramos cinco limitações potenciais de protocolos globais. Concluímos a análise discutindo medidas instrucionais detalhadas, incluindo ferramentas relacionadas a métodos automatizados de observação, como uma alternativa com o potencial de superar várias das limitações fundamentais dos protocolos globais.
Palavras-chave: Aperfeiçoamento de Professores, Métodos de Instrução, Observação, Tecnologia Educacional

Introduction

Teacher observation protocols are a central element of efforts to evaluate teachers and inform instructional improvement efforts.[1] An essential feature of many modern protocols is their "global" approach to measuring instruction. Global protocols provide a summary evaluation of multiple domains of instruction from observers' overall review of classroom processes, by for example, scoring features of classroom discourse over the entire class session or interval of class time observed. Such protocols enable an observational approach to studying classroom instruction that offers both a powerful lens on the distribution of opportunity to learn between and within schools, as well as a framework for instructional improvement grounded in knowledge and information about the teaching process. Yet, recent evidence suggests it is often difficult to obtain a robust portrait of classroom talk and other instructional features using existing protocols, both for research purposes and in applied school improvement efforts (Bell et al., 2014; Gitomer et al., 2014).

In this analysis we argue that both the strengths and limitations of existing protocols stem from their ambitious design parameters; the comprehensiveness and broad intended use of these protocols come with tradeoffs. Empirically, we demonstrate these concerns with reference to specific elements of each protocol and features of their use in the Measures of Effective Teaching Study. Along with the Understanding Teaching Quality (UTQ) study and other research since approximately 2010, the Measures of Effective Teaching study represents an unprecedented effort to understand the limits and possibilities of teacher observation using a suite of well-developed observational protocols that are most commonly used in school districts. Analyzing the MET data, we hypothesize that global protocols have several basic limitations that limit their potential to inform both practice and research, posing and answering the following research questions:

(RQ1) Do global protocols offer precise discriminations in lesson quality?
(RQ2) How independent are protocol sub-domains that are designed to capture different aspects of effective teaching?
(RQ3) How sensitive is measurement reliability to rater training?
(RQ4) Do protocols identify the teacher's own contribution to instructional quality beyond what students themselves may bring to the classroom?
(RQ5) Do protocols exclusively evaluate a continuum of effective practice, making it difficult or impossible to detect tradeoffs and instructional adaptation?

As a set of five research questions, we investigate the extent to which these hypothesized limitations of global protocols are present in the MET data. Some of these concerns have been discussed with regularity in the literature (#3; e.g. Bell et al., 2014; White, 2018), others became especially salient to us in our own analyses seeking to document interactions between instructional practices and classroom composition in the MET data (Aucejo et al., 2018), an investigation relevant to understanding the effects of between and within school sorting of students as well as teacher evaluation policies. In particular, concern #4 and #5 pose great difficulty in understanding how teachers might adapt instruction to match the needs of students.

[1] This research was supported by a grant from the Institute of Education Sciences (R305A170269). Any opinions, findings, and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of IES.

We conclude by discussing how the development of fine-grained observational tools—for example, ones that record and carefully analyze individual utterances, questions, turns at talk, etc.—offer the potential for exceptionally reliable and precise teacher assessment and feedback and might prove an important complement to traditional teacher observation, overcoming the limitations we highlight.

Teacher Observation, Learning, and Instructional Growth

The use of classroom observation to drive school improvement is situated within both accountability and process models of school improvement. In this analysis we focus on the utility of teacher observation within the latter conceptual framework that views teacher observation as providing a critical external source of information to spur teacher learning and instructional growth (Clarke & Hollingworth, 2002; Goe et al., 2012).

At the end of the NCLB reform era, there is an increasing consensus that process models of improvement, including evidence-based innovations, rather than staffing, accountability, or incentive systems, hold the most promise for improving teachers' practice (Gamoran, 2012; Gamoran, 2013; Kelly, 2012). For example, in their 2011 NRC report on test-based accountability, Hout and Elliot conclude, "…overall effects on achievement tend to be small and are effectively zero for a number of programs." (National Research Council, 2011). Similarly, in their evaluation of three randomized trials of incentive pay programs for teachers, Yuan et al. (2013) report that teachers "...did not report their program as motivating," and "…none of the three programs changed teachers' instruction." In contrast, a variety of reform approaches more closely related to the process of instruction, from comprehensive school reform models (see e.g. Borman et al., 2007), to professional development/coaching (Biancarosa et al., 2010) to one-on-one tutoring (Farkas & Durham, 2007) have proven effective in improving the quality of instruction and achievement growth.

Observational information may greatly enhance capacity for instructional improvement in two ways. First, teacher learning entails growth in pedagogically relevant knowledge (Ball et al., 2008; Shulman, 1987) and the ability to evaluate classroom processes and outcomes in light of that knowledge. Classroom observations provide an external viewpoint into classroom processes and outcomes, and an opportunity to structure evaluation around core concepts of teaching and learning. Relatedly, observational protocols can be used to link pre-service teacher education to in-service teacher professional development, generating continuity in teacher learning by carrying forward concepts and concerns. Second, teacher learning may often occur when teachers make improvements through experimenting with classroom practice and then reflecting on outcomes (Clarke & Hollingworth, 2002; Goldsmith et al., 2014). Observational systems can provide structure, motivation, and feedback in the process of "experimentation," reflection, and evaluation by peers, mentors, and teachers themselves.

Overall, instructional reform requires professional development (PD) that helps teachers build and use new knowledge and pedagogical approaches (Darling-Hammond & McLaughlin, 1995). While the professional development that teachers have traditionally received has been criticized as intellectually superficial (e.g. Ball & Cohen, 1999), when PD efforts are substantial, and founded on scientifically-based research (as in the case of university-district partnerships), they are in many cases highly effective (Yoon et al., 2007). Observation systems can potentially facilitate a wide-array of process interventions with broad application to *job-embedded* teacher professional development (Camburn, 2010; Camburn & Han, 2015; Croft et al., 2010; Desimone et al., 2002; Putnam & Borko, 2000), including activities that strengthen social ties among teachers in a professional learning community (Coburn & Russell, 2008; Darling-Hammond et al., 2009; Penuel et al., 2009).

An Overview of Global Teacher Observation Protocols

The use of teacher observation protocols by districts and states to evaluate teachers and inform practice is now well established. Prompted by the federal Race to the Top initiative (RttT), many states (e.g. New York, Ohio, Tennessee, Ohio, North Carolina, Colorado, Michigan) adopted composite systems of teacher evaluation that included systematic teacher observations along with other components. For example, the New Jersey Department of Education's Teacher Evaluation plan, *Achieve NJ*, rates teachers on a four-category scale (highly effective to ineffective), where teacher practice on a state-approved observation instrument accounts for 70-85% of the total evaluation score.[2] New Jersey allows individual districts to choose from a wide-array of protocols in conducting teacher evaluations, including versions of the Danielson Framework for Teaching or FFT (The Danielson Group, 2013), the Classroom Assessment Scoring System or CLASS (see e.g. Hamre et al., 2013), Marzano's Causal Teacher Evaluation Model (Marzano Research Laboratory), and others (New Jersey Department of Education, 2019). In the case of New Jersey, teacher evaluations weighted heavily on classroom observation scores serve multiple monitoring and staff improvement purposes: "All teachers receive individual professional development plans based on their ratings. Teachers rated Ineffective or Partially Effective work with their principals to create a Corrective Action Plan with targeted professional development for the subsequent year. To maintain tenure, all teachers (regardless of hire date) have to continue to earn a rating of Effective or Highly Effective" (New Jersey Department of Education, 2019).

In contrast to the simple subjective ratings of teachers' overall performance sometimes used in employment decisions (see Brandt et al. 2007), the global observation protocols encouraged by RttT generally have several properties: (1) they allow trained raters to score observations of teaching on multiple dimensions and sub-domains, often totaling several dozen or more specific elements of teaching, (2) are based on rigorous research and/or are aligned with established teaching standards, and (3) are designed not only for evaluation but to enhance professional development. Relative to simple survey reports used in large-scale research (e.g. Gamoran & Carbonaro, 2002; Kelly & Majerus, 2011; Newmann et al., 1996; Raudenbush et al., 1993), these systems are a great leap forward in offering the opportunity for independent, occasion-specific measures of teaching.

In addition to FFT and CLASS referenced above as approved by New Jersey for use in teacher evaluation, numerous observation systems have been developed including the Mathematical Quality of Instruction or MQI (Hill et al., 2008), the Protocol for Language Arts Teaching Observation or PLATO (Grossman et al., 2014), the Classroom Strategies Assessment System or CSAS (Reddy & Dudek, 2014), The Thoughtful Classroom (Silver Strong & Associates), The Five Dimensions of Teaching and Learning (The University of Washington, Center for Educational Leadership), and the TAP Rubric (National Institute for Excellence in Teaching). Although the 2015 Every Student Succeeds Act gave states greater autonomy over teacher accountability, Close et al. (2018) report most states are still using "the same or slightly different versions of the previously required systems."

In general these protocols provide a comprehensive assessment of instruction that follows a formative (in the statistical sense) conception of instruction insofar as distinct

[2] The *Achieve NJ* system places lower weight on teacher practice observations (70%) if the teacher is in a tested grade/subject) than if not (85%). Non-tenured teachers must receive a minimum of three, 20-minute observations by multiple observers, while tenured teachers must receive a minimum of two observations per year.

components of instruction collectively constitute rather than reflect effective instruction (See Jarvis et al., 2003, for discussion of formative construct indicators). While some are clearly used for evaluative purposes, they are also well suited for informing teacher learning within the context of a variety of instructional leadership practices, mentoring, induction, and other organizational improvement efforts. Importantly, the original intent of some global protocols was explicitly for research and teacher development. For example, in one of their early working papers, the developers of PLATO stated, "Ultimately, we hope to create a tool that is not only useful for research on teaching, but can be used for teacher development as well." (Grossman et al., 2010). Here we briefly provide further information on the specific protocols used in the MET study and what we see as their particular strengths and special features.

Protocols Used in MET

The Danielson Framework for Teaching or FFT, first designed in 1996, (The Danielson Group, 2013) is a comprehensive observational instrument designed to apply to all disciplines and a wide array of grade levels. FFT includes four domains: (1) planning and preparation, (2) the classroom environment, (3) instruction, and (4) professional responsibilities, with Two and Three pertaining to in-class observation, and One and Four entailing additional materials and out-of-class interaction with the teacher. Within domains Two and Three, in the MET study, classroom instruction was scored on a total of eight components (sub-domains) on a four-point scale: unsatisfactory, basic, proficient, or distinguished. To structure scoring, components are further comprised of elements and raters look for indicators and critical attributes of performance at a given level. A set of possible examples for each component aid in assigning a score. In our view, FFT succeeds in offering an exceptionally comprehensive and broadly applicable observational protocol (like CLASS it can be used for a wide variety of subject matter areas and grade levels but is less explicitly focused on teacher-student interactions; Goe et al., 2012). In addition, while the framework is intentionally engagement-focused and student-centered (The Danielson Group, 2013, p. 5), it is a well-balanced protocol with emphasis on challenge and content coverage.

The Classroom Assessment Scoring System or CLASS (Hamre et al., 2013; La Paro et al., 2004) is a standardized observational system that focuses in particular on the quality of teacher-student interactions. CLASS is organized into three domains (emotional support, classroom organization, and instructional support), each having several subdomains further defined by multiple indicators. Several versions of CLASS are now available that offer a tailored system sensitive to the developmental and pedagogical context of students at different grade levels. In MET, raters scored 15-minute intervals of classroom instruction at the dimension level using a 7-point scale labeled simply from low to high. While there is overlap in the constructs measured by FFT, CLASS has an especially strong focus on emotional support in defining that at the domain level. Hamre et al. (2013, p. 466) describe CLASS's emphasis on social and emotional supports as targeting "key elements" of instruction, which is well motivated by developmental theories of self-determination and attachment. At the same time, we view it as a well-balanced protocol, with for example, essential indicators of challenge within the dimensions of instructional support (e.g. concept development, analysis and problem solving).

The Protocol for Language Arts Teaching Observation (PLATO) is a classroom observation tool designed for 4-9th grade English/Language Arts (ELA) instruction (Grossman et al., 2014). PLATO focuses on 13 elements of instruction related to four underlying domains of instruction: the disciplinary demand of classroom activity and discourse, instructional scaffolding of ELA content, representations and use of content, and the classroom environment. In MET, 8 of the 13 elements were scored during 15-minute instructional segments on a 4-point scale (almost no evidence, limited evidence, evidence with some weaknesses, consistent strong evidence). While there is conceptual

overlap with other observational protocols (e.g. with CLASS's instructional support domain), the PLATO elements have their origin specifically in research on English/language arts instruction and are closely linked with literacy learning (e.g. the text-based instruction element). Early research with PLATO found that the choice of instructional activity (e.g. whole class literature instruction vs. small group or paired writing instruction) affected PLATO scores; as such, PLATO is most reliable when aggregated over several lessons that capture a range of instructional activity (Cor, 2011; Grossman et al., 2010).

The Mathematical Quality of Instruction (MQI) protocol was developed by Heather Hill and colleagues as a subject-specific observational system between 2003 and 2010 (Hill et al., 2008; Hill et al., 2011). MQI considers 6 elements (richness of the mathematics; errors and imprecision; working with students and mathematics; student participation in meaning-making and reasoning; explicitness and thoroughness; and connections between classroom work and mathematics) that concern the relationships between teachers, students, and content.[3] In MET, raters scored observations with lessons broken into 4–7.5 minute segments, as well as the whole observation, on a 3-point scale (dimension not present, partially present, or predominantly present) using scoring anchors provided for each score on each element.[4] The data include both the holistic ratings based on the whole observation and the segment-level ratings. As with PLATO, even though it is subject-specific there is conceptual overlap with MQI and the other protocols (e.g. student participation in meaning-making captures constructivist principles that align with the FFT protocol). However, MQI stands out as being the most heavily focused on teachers' exhibition of pedagogical content knowledge in mathematics. Indeed, in MET, the MQI raters also provided a "lesson-based guess" at mathematical knowledge for teaching.

Data and Methods

The Measures of Effective Teaching Study collected data on teachers' instructional practices in six school districts over a two-year period from 2009-2010 to 2010-2011. Participating school districts were generally very large districts, encompassing but not limited to urban central-city schools in all cases: Charlotte-Mecklenburg (NC) Schools, Dallas (TX) Independent School District, Denver (CO) Public Schools, Hillsborough County (FL) Public Schools, Memphis (TN) City Schools, and the New York City (NY) Department of Education. As general indicators of socio-demographic context, MET teachers ranged from a low of 21.8% white in Memphis to a high of 92.2% in Denver (56.8% overall), while 34% of MET students were white and 54% received a subsidized school lunch (Kane et al., 2012, pp. 16-17).

While the random assignment of teachers to classrooms in Year 2 and other important features of MET are described elsewhere, here we provide a brief overview of relevant features of the teacher sample and observation process (See also the MET user guide, ICPSR document 34771). The MET study scored video observations of instruction using the high-quality observational protocols previously described. In contrast to typical use in evaluation, there were no stakes attached to the observational measures in MET, they were collected purely for research purposes. Also unlike in typical use, the raters were not local school administrators, curriculum coordinators, or lead

[3] Teachers are also rated on Explicitness and Thoroughness, but only for about 8% of the observations in the MET data.

[4] An exception is that "Classroom Work Connected to Mathematics" in MQI is measured on a binary scale. As a minimal indicator of mathematical content, it may be useful given that MQI segments were only 7.5 minutes long in the MET study, but it is not a feature found in other protocols so we do not include this in our tables.

teachers in the teachers' own schools, but impartial expert raters. While we will concentrate on analyzing the limitations of these protocols here, they were state of the art measures at the time of the MET data collection.

Within each of the six MET districts, teachers were voluntarily recruited from "traditional" public elementary and middle schools; 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).[5] Participating teachers who met a few basic eligibility parameters (e.g. 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) received a $1500 incentive. Overall, MET included a diverse sample of teachers (only 56.8% were white), generally representative of their districts (Kane et al., 2012).

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. In Year 1, approximately 1570 teachers contributed videos for CLASS and FFT scoring and 940 teachers contributed videos for MQI and PLATO scoring. The number of videos also varied by observation protocol with roughly 7800 for CLASS and FFT and 3500 for PLATO/MQI. Finally, some protocols have fewer ratings per video; for example FFT has about one rating per video while CLASS has closer to two (sample sizes in the first two tables report the number of ratings for each observation protocol).

Observation and 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. The recorded lessons were 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 a single video/audio channel 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 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.[6]

Analytic Plan

To assess the extent of the five postulated limitations of global observation protocols, we perform descriptive and inferential analyses across multiple levels of analysis, including ratings of the same classroom observations across different observers, features of observation

[5] Schools were recruited using a combination of small monetary and equipment incentives (the video equipment used in the study was donated to each participating school).

[6] Raters did not code any videos from districts in which they worked.

scores across lessons of the same teacher and in the sample as a whole, and observation scores in relation to features of teachers and classrooms. While analyses related to some of the questions can be found in the existing literature, with the exception of Table 3, we provide novel analytic content in our calculations and comparisons.[7] Although the sample varies somewhat as noted, we generally rely on the Year 1 data, for grades 4-9, when teacher placement in classrooms occurred naturally. In Table 4 we include both Year 1 and Year 2 data (teacher assignment in Year 2 is random rather than naturalistic), because we seek to make inferences about potentially small but meaningful compositional effects where statistical power is a major concern.

Research Question One: Discriminations in lesson quality. To answer our first research question we report observation-level descriptive statistics for the four protocols. To measure the extent to which the ratings discriminate between lessons of different quality, we consider standard deviations of each subdomain and the number of scores that cluster at the modal rating categories (there is generally a sharp drop off between the prevalence rate in the modal categories and the remaining categories). For example, FFT has four response categories, with the middle two categories representing a strong mode. We also report skewness of the distribution (positive values indicating the mean exceeds the median), and the standardized kurtosis (values greater than 3 are more peaky than the standard normal distribution).

Research Question Two: Independence in sub-domains. To answer our second question, we perform a basic exploratory factor analysis at the level of the subdomains within protocols to consider whether there is evidence of the subdomains being independent enough to pick up different aspects of teaching, or *factors*. The metrics we use include several often used rules of thumb for construct independence: the number of factors with eigenvalues greater than 1 (*Kaiser criterion*), the number of factors needed to explain 90% of the variance in the covariance structure (eigenvalues), and the cumulative proportion of variance explained by adding each additional factor (Richards et al., 2013).

Research Question Three: Reliability and training effects. While Question One and Two rely on observation-level data (a unique observation for each class session as in real-world use), question Three relies on a given class session being scored by multiple raters, which occurred as part of the MET's inter-rater reliability investigation (a phase of analysis commonly done early in a study before the rating process occurs at scale). For each of the four protocols, we compare three measures of inter-rater reliability across domains (percent exact agreement, simple kappa, and quadratic weighted kappas). These measures are reported using two different sets of rater pairs; rater pairs where both were trained normally and rater pairs where one rater was an expert coder. The comparison across rater types is a rough indication of the importance of coder expertise and training.

Percent exact agreement statistics, while intuitive, are difficult to compare across protocols, because as the number of response categories increases, the likelihood of exact agreement falls accordingly (e.g. for CLASS which has 7 categories). Simple Kappa statistics take into account base-rate chances of agreement, making them well-suited for comparing reliability across protocols with different base rates, and especially different sub-domains within a given

[7] For example, the results in Table 4 build closely on Campbell and Ronfeldt (2018), but we also report results for CLASS, PLATO, and different forms of MQI. Even Table 3 is novel insofar as we provide an explicit side-by-side comparison of non-expert and expert ratings that highlights sensitivity to training in a way not discussed in the official MET reports.

protocol (which generally have the same number of response categories). Finally, quadratic-weighted Kappa statistics additionally adjust for the number of response categories (i.e. ratings in near vs. far categories are evidence of consistency), making them more comparable across protocols and providing the most holistic comparison to the mathematical model of independence (akin to a chi square statistic).[8]

Research Question Four: The teachers' own contribution to instruction. For the fourth question, we provide indirect evidence on how sensitive the MET protocols might have been to influences of social norms and process beyond the teachers' control. We build closely off of Campbell and Ronfeldt's (2018) analysis of FFT to analyze the association between the teacher rating from each of the protocols, measured by averaging scores across subdomains, and classroom characteristics, including average prior test performance on ELA and math, percentage black, Hispanic, Asian, other non-white, male, special education, English language learners, free/reduced-price lunch, class size and teacher value-added. The extent to which classroom characteristics are individually or jointly significant predictors of FFT, CLASS, MQI or PLATO may be suggestive of influences beyond the teacher's control. We also consider the summative significance by considering the total explanatory power. Controlling for teacher value-added helps to control for the alternative explanation that certain classroom characteristics are being matched to more effective teachers.

Research Question Five: A continuum of effective practice. Are sub-domain scores always positively correlated? For question 5, we examined the covariance matrix of the sub-domains of the MET protocols. A weak test of this question would rely on the covariance matrix for the data as a whole. However, in this case consistently positive pair-wise correlations (e.g. domain 1 of FFT paired with domain 2, then with domain 3, etc.) are likely to be found because teacher training, effort, etc. induce a positive association. Instead, we focus on intra-teacher variance (i.e. looking only within class sections, a teacher and specific class of students, and pooling the results). Thus, we consider how often negative pairwise correlations occur for specific sections/teachers, which might be indicative of potential tradeoffs that teachers may face in choosing emphasis among the subdomains. Given that we have about 8 observations per class section (varying across protocols), some amount of weak negative correlations will occur merely due to chance. Thus, we count the number of pair-wise correlations that was (negatively) greater than −0.2 (a somewhat arbitrary cut-point, but it should remove some of the chance negative correlations). For example, for FFT, we consider 28 pair-wise correlations among 7820 observations nested within 921 sections. We present the proportion of negative pair-wise correlations, averaging across sub-domains, along with the sub-domain pairing with the highest average (most negative instances) and lowest average (least negative instances) incidence rates of apparent tradeoffs.

Results

Imprecise Discriminations in Lesson Quality

While the comprehensive focus of existing global observational protocols is noteworthy, one risk of trying to capture so many dimensions of teaching may be that only rough, imprecise distinctions can be made concerning specific domains of instruction. To answer our first research question, Table 1 reports observation-level descriptive statistics for the four protocols,

[8] Agreement statistics such as Gwet's AC (Gwet, 2008) are further useful if base rates of agreement are very high. As that is not an issue in the present study we report the more commonly used Kappa statistics.

Table 1
Protocol Score Summary Statistics by Observation Instrument Components[a]

	Proportion of observations in the Modal Categories/ scores[b]	**Modal Domain Scores Used**	**Mean**	***SD***	**Skewness**	**Kurtosis**	***N***
FFT							
Using Questioning and Discussion Techniques	0.87	2,3	2.20	0.64	0.05	2.82	7820
Communicating with Students	0.96	2,3	2.61	0.57	−0.35	2.70	7820
CLASS							
Positive Climate	0.73	4,5,6	4.44	1.28	−0.17	2.50	17804
Quality of Feedback	0.73	2,3,4	3.55	1.28	0.19	2.57	17804
Instructional Dialogue	0.76	2,3,4	3.28	1.28	0.30	2.57	17804
Lack of a Negative Climate	0.98	5,6,7	1.42	0.77	2.47	11.06	17804
PLATO							
Strategy Use in Instruction	0.81	1,2	1.73	0.88	0.99	3.04	7546
Behavior Management	0.93	3,4	3.70	0.65	−2.39	8.37	7546
MQI							
Overall	0.82	2.00	1.93	0.41	−0.49	5.54	14124
Lack of Errors and Imprecision (Holistic)[c]	0.79	3.00	2.75	0.50	−1.94	5.93	14124
Explicitness and Thoroughness (Holistic)[c]	0.58	2.00	1.72	0.59	0.16	2.45	2500
Student Participation in Meaning Making and Reasoning (Holistic)[c]	0.81	1.00	1.20	0.42	1.86	5.36	14124
Lack of Errors and Imprecision	0.81	3.00	2.78	0.48	−2.07	6.54	14124
Explicitness and Thoroughness	0.48	1.00	1.60	0.63	0.59	2.39	2489
Student Participation in Meaning Making and Reasoning	0.82	1.00	1.19	0.42	2.06	6.42	14124

[a] The protocol components were chosen by largest and smallest standard deviation.
[b] The modal score categories depend on the number of score categories for each instrument. FFT has four, CLASS has seven, PLATO has four and MQI has three.
[c] Raters performed scoring using the MQI protocol on segments of a lesson (Holistic label excluded) but also gave a score to the whole lesson (Holistic label included).

highlighting the sub-domains with the smallest and largest standard deviations. As a representation of the tendency for observations to cluster in one or a few categories (yielding a low standard deviation), we report the proportion of observations that fall into adjacent modal categories.[9] For example, on the 8 FFT sub-domains, the proportion of teacher observations in the middle two out of four categories (basic, proficient) ranged from a low of 87% (Using Question and Discussion Techniques) to a high of 96% (Communicating with Students). For FFT, the sub-domains exhibit little variation in their distributions with highly similar standard deviations, and modest differences in skewness and kurtosis. PLATO is also scored on a four-category scale, and shows somewhat greater spread than FFT. Yet, even the sub-domain of PLATO with the most variable scores (Strategy Use and Instruction), has 81% of observations in the bottom two categories.

CLASS is scored on a 7-point scale anchored with "low," "middle," and "high," and thus offers the possibility of greater distinctions. Yet, even the most variable sub-domains (three are tied with a SD of 1.28) contain 73% or more of the observations in just 3 out of 7 categories, while the least variable (lack of negative climate) contains 98% of the observations in three categories (and 92% in the top two categories). For MQI, we show both the holistic scores and the individual 7.5-minute segment scores. The sub-domain with the greatest variability in both segment and holistic scoring, Explicitness and Thoroughness, contained 48% and 58% respectively in the bottom and middle category though this was only evaluated for a small proportion of observations. The sub-domain with the least variability, Student Participation in Meaning Making and Reasoning, contained 82% and 81% of observations in the lowest category in the segment and holistic ratings respectively. We also report results for the holistic Overall Mathematical Quality of Instruction score, where 82% of lessons score in the middle category. In both the CLASS and MQI protocols, there are fairly substantial differences across sub-domains, some are quite "peaky" while others are less so.

Overall, while it may be the case that in fact a strong majority of teachers' lesson qualities are indeed generally adequate and "in the middle," we worry that the tendency for scores to cluster in a few modal categories in global protocols limits their ability to guide improvement. Further, teachers may disregard protocol results when they do not have examples of high or low ratings from which they can learn how to improve their teaching. They may also become discouraged about teaching when they only rarely see the potential to excel and lack positive feedback from these ratings (e.g. in the case where 5% or less of observations achieve an exemplary score).

Lack of Independence in Sub-Domains

Do global observational protocols capture multiple domains of instructional practice, such that teachers can receive targeted feedback on what areas of instruction most need improvement? In the systems of observation used in MET and the Understanding Teaching Quality (UTQ) study, individual domains of instruction, ostensibly critical to focusing improvement efforts, are not as separable in practice as they are in theory. Liu et al. (2019) examined the covariance structure of FFT observation scores in three sets of data, low-stakes observations from the research-focused UTQ study, and two practice-based implementations in the Understanding Consequential Assessment Systems for Teachers study. In all cases, they

[9] For FFT, the middle 2 out of 4 categories; for CLASS, the modal 3 out of 7 categories, which varies among categories depending on the skewness of the sub-domain; for PLATO, 2 out of 4 categories, which varies by skewness; for MQI 1 out of 3, which is either the bottom or middle category depending on skeweness.

found high correlations across the four FFT domains and eight sub-domains such that a single factor structure best fit the data.

In Table 2, we report results of a basic exploratory factor analysis for all four observational protocols (with multiple specifications of MQI). In all cases, whether one focuses on the number of eigenvalues above 1, the number needed to reach 90%, or the apparent point of diminishing returns as additional factors are added, it is possible to specify a simpler structure, with fewer discernable latent factors than the number of sub-domains that the protocols consist of and intend to measure separately.

Table 2
Factor Analysis of Observation Instruments

Observation Instrument	# of Components Used	Eigenvalues > 1	Factors to Explain > 90% of Variability	R^2 with X Factors 1	2	3	4	*N*
FFT	8	1	2	0.800	0.932	0.960	0.978	7819
CLASS	12	2	4	0.678	0.851	0.899	0.931	17804
PLATO	6	1	3	0.575	0.797	0.991	0.998	7546
MQI HOL [a]	5	1	2	0.840	0.928	0.983	1.000	11509
MQI Non-HOL [a]	5	1	3	0.725	0.881	0.946	1.000	11509
MQI HOL [b]	4	1	1	0.903	0.991	1.000	1.000	14124
MQI Non-HOL [b]	4	0	2	0.876	0.964	1.000	1.000	14124

[a] Within the MQI instrument HOL refers to a rating of the whole lesson while Non-HOL is at the segment level (each lesson was divided into 7.5 minute segments). The MQI component Explicitness and Thoroughness is excluded since it greatly reduces the sample.
[b] This is the same MQI instrument but the component Classroom Work Connected to Mathematics is excluded as this is rated on a different scale. Additionally, MQI component Explicitness and Thoroughness is also excluded since it greatly reduces the sample size.

One possible explanation for the consistency in sub-domains scores is that training and other teacher quality factors create a similar level of competence across domains. Another source of consistency is that while the sub-domains are discreetly defined, they are also in some cases conceptually similar and part of a larger construct (e.g. in MQI Richness of the Mathematics and Errors and Imprecision both relate to teachers' content and pedagogical content knowledge). An alternative explanation is that features of the observation system, such as a tendency for overall perceptions to create a halo-effect, create artificial consistency in sub-domain scores (Liu et al., 2019; McCaffrey et al., 2015). Humphry and Heldsinger (2014) argue that consistency in the structural design of rubrics (global teacher observation protocols fall into the general category of a rubric), where all scoring domains have the same small number of response categories, can create similarity in domain scores both because raters are prevented from making the finer distinctions they are capable of for some domains, but also because the common structure can generate conceptual overlap and repetition/redundancy in score descriptions. While the protocols we analyze seem susceptible to the structural concern Humphry and Heldsinger describe, actual conceptual overlap, and/or rater focus on over-

arching concepts (e.g. student-centered instruction) rather than discrete domains of teacher practice would remain a concern even with structural modifications.

Reliable Measurement Requires Training and Monitoring

How dependent on expert training are current global protocols in offering reliable assessments of instruction? Certainly, in using any observational data to make inferences about teacher effectiveness, a robust sampling process, with multiple representative observations per teacher is needed. In much prior research, four observations per teacher has been used as a target in making inferences at the teacher level (e.g. Gamoran et al., 1995; Kane et al., 2012; Kelly, 2007). The reliability of measured instruction improves considerably, from, for example, two to four observations (Kane et al., 2012, Table 11), but thereafter begins to reach a point of diminishing returns (Kelly et al., 2018, endnote 8). At a more basic level, affecting reliability at both the teacher and observation level, robust observational measurement requires adequate training and monitoring of observers. Results of observational studies to date suggest that achieving reliable observational ratings of teaching quality is challenging (Bell et al., 2014).

One way to demonstrate the importance of training and rater competency and concentration is to compare the inter-rater reliability under "normal" or at-scale conditions, to inter-rater reliability in a university or research-center setting (in MET, scores from content experts at the research firm contracted to collect the data are available). Table 3 reports inter-rater agreement statistics from the MET data. For each of the four protocols, the domain with the smallest and largest discrepancy in the inter-rater reliability (simple kappa) between rater pairs where both were trained normally and rater pairs where one rater was an expert coder are shown. While rates of exact agreement are often above 70% for the domains in Table 3, agreement in the kappa metric is sometimes not much above chance. Low kappa statistics here reveal that when there are only a few categories to choose from, and some categories are rarely selected by any rater, 70-75% agreement is not particularly impressive.

For the FFT protocol, where 90% of the time raters were adjudicating between just one of two middle categories, basic and proficient, exact agreement by two local raters ranged from only 47.3% to 65.8% across domains, and simple kappas ranged from .05 to .28 (or .21-.45 quad weighted). However, reliability improves substantially for "back-scored" videos, where one of the raters was more extensively trained; the simple kappa's jump to .39-.48, with about 70% exact agreement. Other protocols show similarly large increases in reliability when expert raters are used on some domains.

Overall, both the statistics in Table 3 and results published elsewhere show that in practice the reliability of classroom observations is quite variable, depends on adequate training and monitoring, and at the low end is problematic (Bell et al., 2014; Cash et al., 2012; Cohen & Gitomer et al., 2014; Goldhaber, 2016; Kane et al., 2012). White's (2018) analyses of the UTQ data suggest that current standards for rater accuracy and consistency may be too low. However, it would not be appropriate to label these protocols simply as "unreliable," as levels of agreement among expert raters are well above chance/independence (see quadratic weighted kappas) despite the inherent complexity of the phenomena being rated.

Table 3

Comparison of [normal, normal] and [normal, expert] rater agreement in four Observational Protocol: Summary of domains with smallest and largest discrepancy in the simple kappa between rater conditions[a]

	[Normal, Normal] Rater Pairs[b]			[Normal, Expert] Rater Pairs		
	Segments Scored	% Exact Agreement	Kappa Simple, Quad wtd.	Segments Scored	% Exact Agreement	Kappa Simple, Quad wtd.
FFT						
Managing Classroom Procedures	374	61.5%	.24, .34	2509	69.4%	.41, .58
Using Assessment in Instruction	374	47.3%	.05, .21	2509	66.9%	.39, .50
CLASS						
Behavior Management	587	42.6%	.17, .48	2444	46.9%	.22, .61
Student Engagement	587	27.6%	.04, .36	2444	42.0%	.21, .53
PLATO[c]						
Time Management	182	73.6%	.34, .56	1045	77.0%	.41, .49
Intellectual Challenge	182	44.5%	.12, .33	1045	63.7%	.44, .60
MQI[d]						
Working with Students & Mathematics	171	70.2%	.30, .32	943	75.0%	.38, .39
Errors & Imprecision	171	70.2%	.14, .12	943	80.0%	.47, .55
Overall Mathematical Quality of Inst.	171	73.1%	.05, .10	943	79.5%	.34, .39
L-B-G at Math. Knowledge for teaching	171	76.0%	.18, .28	943	77.7%	.27, .34

[a] Statistics collated from the Measures of Effective Teaching Study Observations Measures Report, Year one data, Phase Two scoring.

[b] In the Observations Measures Report [normal, normal] pairs are referred to as "double scored" while [normal, expert] pairs are referred to as "back-scored"

[c] We report segment 2 statistics for PLATO. Classroom Discourse was tied with Intellectual Challenge as the most discrepent, a difference of .32 between the simple kappas in the double vs. back scored ratings.

[d] MQI statistics are the holistic ratings, segment specific ratings not considered. For MQI we also report statistics for two entirely global ratings, including the Lesson Based Guess at Mathematical Knowledge for Teaching.

Identifying the Teachers' Own Contribution to Instruction

Do global observation protocols primarily reflect the teachers' own contribution to instruction, or are they heavily impacted by features of the learning environment beyond their control? Given their widespread use in evaluation, ideally, global protocols would carefully distinguish between "teacher moves," the teachers' own contribution to instruction, with the enacted quality of instruction influenced by social norms and processes beyond the teachers' control. Consider for example the FFT sub-domain, "creating an environment of respect and rapport." Rubric examples for the proficient category include: "teacher greets students by name as they enter the class or during the lesson"—more obviously a teacher move, as well as "students attend fully to what the teacher is saying"—which is less obviously related to teacher moves. A focus on the enacted quality of instruction may be desirable in assessing instruction and instructional growth toward a target level in an applied setting but may hinder causal research on teacher effectiveness.

To provide indirect evidence on how sensitive the MET protocols might have been to influences of social norms and process beyond the teachers' control, Table 4 reports regression models showing the association between compositional features of the classroom and protocol scores. As in Campbell and Ronfeldt (2018), we find that classroom achievement composition, racial composition, percentage free lunch, and even percentage male are associated with overall FFT scores (e.g. a coefficient of .091 for class mean achievement in our analyses is the same to two significant digits as estimated by Campbell and Ronfeldt). Classroom composition measures are jointly significant predictors of FFT and CLASS at the 99% confidence level, but this is not the case for MQI or PLATO. Yet, examining the change in R^2 from the model with and without classroom controls, the additional explanatory power from including the rich classroom composition measures is quite small, at most 0.003. Considering the magnitude of estimated effects of average classroom initial achievement, teachers who have 1 standard deviation higher average classroom achievement will be rated 0.09 of a standard deviation higher on FFT and 0.07 on CLASS. It remains an open question whether the magnitude of these effects would substantially contaminate either teacher evaluation, evaluations of curricular reform, etc., but does raise a note of caution particularly when comparing teachers with very different student compositions in high stakes settings. We also note that the correlations with classroom composition could reflect teacher adaptation, and thus a potentially productive aspect of teaching, though we cannot distinguish a measurement problem from adaptation in these protocols.

Table 4
Regressions Showing Sensitivity of Observation Protocols to Classroom Composition [a]

Classroom Characteristics [b]	**(1) FFT**	**(2) CLASS**	**(3) PLATO**	**(4) MQIHOL** [c]	**(5) MQI nonHOL** [c]
% Black	−0.244	−0.119	−0.248	0.062	0.112
	(0.171)	(0.101)	(0.316)	(0.213)	(0.162)
% Hispanic	−0.216	−0.005	−0.231	0.286	0.196
	(0.158)	(0.098)	(0.308)	(0.186)	(0.136)
% Asian	−0.418*	−0.204	0.035	0.020	0.193
	(0.244)	(0.150)	(0.532)	(0.298)	(0.196)
% Other nonwhite	−0.766**	−0.242	−1.374	0.100	0.045
	(0.303)	(0.199)	(0.869)	(0.393)	(0.264)
% Male	−0.158*	−0.110*	−0.223	−0.197	−0.087
	(0.088)	(0.065)	(0.267)	(0.121)	(0.084)
% SPED	−0.034	−0.024	0.394	0.036	0.088
	(0.122)	(0.075)	(0.270)	(0.155)	(0.114)
% ELL	0.079	−0.008	−0.026	0.003	−0.039
	(0.121)	(0.075)	(0.281)	(0.128)	(0.089)
%FRPL	−0.143	−0.026	0.343	−0.100	−0.036
	(0.094)	(0.058)	(0.233)	(0.121)	(0.080)
Avg prior test performance	0.091**	0.074***	0.016	0.047	0.029
	(0.035)	(0.023)	(0.073)	(0.047)	(0.035)
Class size	−0.003	−0.004**	−0.008	−0.001	−0.001
	(0.003)	(0.002)	(0.006)	(0.003)	(0.002)

Table 4 cont.

Regressions Showing Sensitivity of Observation Protocols to Classroom Composition [a]

Classroom Characteristics [b]	(1) **FFT**	(2) **CLASS**	(3) **PLATO**	(4) **MQIHOL** [c]	(5) **MQI nonHOL** [c]
Teacher value-added	0.159***	0.035	0.112	−0.015	−0.007
	(0.055)	(0.034)	(0.150)	(0.077)	(0.058)
N	9686	20875	5526	17524	17524
R^2	0.378	0.295	0.371	0.221	0.127
R^2 without classroom composition variables	0.375	0.293	0.369	0.219	0.127
Joint significance test (classroom composition variables)					
F	2.75	3.27	0.98	0.72	0.59
p-value	0.0024	0.0003	0.4601	0.7099	0.8242

Notes: *** denotes significance at the 1% level, ** at the 5% level and * at the 10% level. Standard errors are clustered at the teacher level. Regressions also control for indicators for year, grade and whether the lesson was an ELA lesson in the case of CLASS and FFT.

[a] The dependent variable for these regression is an average of the scores for each protocols components.

[b] These variables starting with black students and ending with FRPL students represent proportions at the classroom level.

[c] Within the MQI instrument HOL refers to a rating of the whole lesson while Non-HOL is at the segment level. (Each lesson was divided into 7.5 minutes segments). The MQI components Classroom Work Connected to Mathematics and Explicitness and Thoroughness are excluded since they greatly reduce the sample.

A Focus on a Continuum of Effective Practice

Are existing global observational protocols only effective at analyzing instruction along a continuum of effective practice? While a focus on effective practice aids use in evaluation, it may hinder basic research on instruction and learning which involves tradeoffs in terms of how time is spent and in emphasis. In contrast, many fine-grained approaches to classroom observation, as well as measures of assignment quality, measure instruction more agnostically. For example, in Nystrand and Gamoran's program of research on ELA instruction, classroom time-use is exhaustively coded, but no a-priori judgement is made about the most appropriate ratio of say, small group work to whole-class instruction (Nystrand & Gamoran, 1997). Likewise, assignment quality protocols like the Intellectual Demand Assignment Protocol, do not privilege particular teaching practices, or do so less inherently (Joyce et al., 2018; Wenzel et al., 2002). Thus, in such systems, it is imminently possible to detect *trade-offs* in instructional practice (as the case of time-use exemplifies).

To investigate the potential difficulty in detecting instructional trade-offs, we examined the covariance matrix of the sub-domains of the MET protocols. At an aggregate level across the Year 1 observations as a whole, all—every single one—of the pair-wise correlations between sub-domains within the protocols are positive, suggesting that teachers who are effective in one domain are generally effective in other domains.[10] However, at the level of a given observation, the protocols may in fact detect trade-offs, with a teacher scoring low on one domain but high on other(s).

Table 5 shows that across all sub-domains, instances (proportions) of pair-wise negative correlations within class sections (a specific teacher and group of students) ranged from .08 (FFT) to .17 (CLASS). Interestingly, for MQI, this is somewhat more likely to occur in segment level scoring

[10] With the exception of MQI's measure of classroom work connected to mathematics, which is a simple binary (yes, no) measure that only minimally captures instructional quality.

Table 5

Proportion of Negative Correlations Overall and Averaged Across Components [a]

Observation Instrument	Overall Proportion of Negative Correlations	Average Proportion Across Domain Components [b]				
		High Average Proportion	High Component	Low Average Proportion	Low Component	Number of sections
FFT	0.08	0.09	Managing Classroom Procedures	0.07	Using Questioning and Discussion Techniques	921
CLASS	0.17	0.21	Behavior Management	0.12	Quality of Feedback	2235
PLATO	0.13	0.15	Intellectual Challenge	0.12	Behavior Management	1204
MQI HOL[c]	0.09	0.06	Lack of Errors and Imprecision	0.04	Lack of Errors and Imprecision	1281
MQI Non-HOL[c]	0.14	0.13	Student Participation in Meaning Making and Reasoning	0.09	Student Participation in Meaning Making and Reasoning	1281

[a] For this analysis a correlation is considered negative if it is < -0.2

[b] Notes: The average proportion across components pairs a single component with every other component from that observation protocol and takes the average of the proportion of negative correlations across pairs.

[c] Raters performed scoring using the MQI protocol on segments of a lesson (Non-HOL) but also gave a score to the whole lesson (HOL).

than in holistic scoring.[11] Although there are several instances in which particular sub-domains are highly unlikely to negatively co-vary with other domains of instructional practice (e.g. those listed in the low column), there are also a number of other domains that routinely, if not typically, negatively co-vary with other domains (e.g. CLASS's Behavior Management domain or PLATO's Intellectual Challenge). Thus, we conclude from this analysis that while detecting trade-offs in instructional practice is clearly not a strength of these protocols (in particular FFT), it does in fact occur in practice.

Discussion

While we have focused on posing and answering questions about the limitations of global protocols for teacher observation it is important to remember the advantages of such tools, which are used in many districts to measure teacher effectiveness. An underlying goal of global observational protocols is to organize and communicate best practices for teachers and those who support them. As such, some of the characteristics we have highlighted, like their comprehensive nature and focus on a continuum of effective practice, are logical design features and give the protocols a wide array of instructional improvement uses. These features may also make the protocols especially useful for some research purposes, helping to move educational research beyond studies of achievement alone to create a richer understanding of opportunity to learn. For example, we have used the MET protocols to document the distribution of opportunity to learn across schools. How disparate is instruction in different schools and are school-to-school differences in instruction associated with students' family background and other compositional features of schools (Kelly et al., 2019)? Comprehensive measures appropriately rooted in the best practices literature are well suited to answering those questions.

The utility of global observational protocols must also be understood in the context of the use of standardized test score data to identify teacher effectiveness. Because simple observable characteristics like degree attainment, certification and even experience do not adequately capture teacher effectiveness, the literature has focused on a fairly low-cost alternative (at least in a regime with annual student testing) of value-added based measures of teacher effectiveness (Gamoran, 2012; Kane et al., 2012). Value-added measures have several well-known limitations (Koedel et al., 2015; Jackson et al., 2014; Stacy et al., 2018), including concerns about imprecision and that they do not provide useful information to help teachers improve their teaching. Combining insights of value-added measures with global observation protocols has the potential to help address both of these concerns and is a practice that is increasingly being adopted in districts. In fact, among the important insights of the MET study was that combining multiple measures of teaching practice, including global observation protocol scores, provides more stable estimates of teacher effectiveness than value-added measures alone (Cantrell and Kane, 2013; Mihaly et al., 2013). The MET study also revealed important insights for increasing reliability of ratings, such as having more than 1 observer per teacher, observing more than 1 lesson and supplementing full lesson observations with short observations (Cantrell & Kane, 2013).

Yet in this study, we sought to document some of the limitations that have emerged in our analysis of the global protocols as used in MET, particularly in their potential to help teachers improve practice and for researchers to answer important questions about effective teaching practice. We find that global protocols provide only imprecise discriminations in lesson quality;

[11] MQI segment scoring started by dividing lessons into 7.5 minute segments. Raters graded each segment and assigned a rating (1-3 scale) for each sub-domain as well as rating for the whole lesson (the holistic score).

some of the distinctions are so rough they may only be useful for guiding instructional reform for a small minority of teachers. In other cases, including examples in MQI and CLASS, greater discriminatory power emerges.

Each of the protocols used in MET are comprised of multiple sub-domains, such that it is possible in theory to identify teachers strong in one area, but not in others, and to provide feedback on specific areas that need improvement. However, analysis of the covariance structure of the sub-domain scores finds that they are not as orthogonal in practice as they appear to be in their construction. It is not clear whether this phenomenon reflects true underlying similarity in teacher competence across domains, or some issue in measurement. If the latter, then this feature limits the protocols' use in formative feedback.

We also summarize a finding evident in analyses of the MET rating process published in the Observation Measures Report; the reliability of the protocol is highly dependent on training and monitoring, and highly variable under the conditions of the MET study. At the low end, levels of agreement are only slightly better than chance. A further concern was that global observation protocols might have difficulty separating the teachers' own contribution to instruction from what students bring to class at the start of the year. This concern seems evident at times in the construction of the protocols themselves, and there is indirect evidence of this possibility in Campbell and Ronfeldt (2018), Steinberg and Garrett (2016), and our own replication of Campbell and Ronfeldt's analysis of the relationship between classroom composition and FFT scores. However, when we examined the larger set of protocols, we found that in fact classroom composition is not always related to protocol scores, and certainly not in a predictable fashion. Thus, in our view, it is possible to design global protocols that adequately capture teachers' own contribution to instruction.

Finally, we raised the concern that many global observation protocols focus only on a continuum of effective practice and cannot readily detect tradeoffs in instructional emphasis that occur as teachers adapt to students. This seems highly evident in the construction of the protocols. However, our analyses searching for examples of instructional trade-offs occurring in the data found rates of negatively correlated sub-domain scores that suggests the protocols can at times document tradeoffs, though it is not a strength of these protocols.

These limitations may combine to make many inferences about important instructional processes difficult. For instance, in our work with the MET data (Aucejo et al., 2019), we studied how the benefits of a given practice vary with the composition of the classroom and how teachers adapt to classroom compositions by adjusting their practice. One of our original intentions was to see if these adaptations in themselves might be a measure of teacher effectiveness (Corno, 2008; Nurmi et al., 2013). Our hypotheses implicitly assume that teachers face tradeoffs in how they spend their time in the classroom, and in choice of curriculum and pedagogy. We anticipated, in theory based on descriptions of subdomains in the different protocols, that underlying aspects of teaching practice, such as student-centered approaches, would be common across protocols and thus separable from other aspects of practice. In reality, we found that subdomains within protocols were not as separable as might be hoped for in examining instructional tradeoffs. We also anticipated being able to exploit multiple measures of teaching practice across multiple years to study teacher adaptivity, but found little systematic adaptation, perhaps because many of the measures confound teacher and classroom moves such that it was not really possible to identify *teacher* adaptations. Ultimately, while we made some useful progress in understanding how instructional effectiveness is moderated by classroom composition and elucidated important associated implications for accountability systems, we gleaned from our experience that it is simply not possible to adequately test many important

hypotheses about instruction, such as the tradeoffs teachers face in adapting instruction to diverse student needs, with many global protocols.

Policy Implications of Unreliability and other Global Protocol Limitations

In Rothstein and Mathis's (2013) response to final reports prepared by the Measures of Effective Teaching Study researchers (as opposed to researchers later conducting secondary data analysis), they argue that findings from MET on the reliability of observational methods and on the relationship between observational scores and value-added, "say little about how best to conduct teacher evaluations in the real world." In this analysis, we have taken a more expansive view of observational protocols, which can be used for evaluation, but also professional development and research purposes.

Depending on the use case, some of the concerns we have outlined here present more serious policy implications than others. Problems of sub-domain independence (limitation #2) and focusing only on a continuum of effective practice (#5) are grave concerns for research. In contrast, imprecise discrimination in lesson quality (#1) is a major problem for use in professional development activities. Limitations 1, 3, & 4 are all potentially problematic for use in evaluation, although limitations 1 and 3 seem most serious in these data. However, at this time we believe that educational professionals setting and implementing policy should not make decisions about the role of observations in teacher evaluation on the basis of these limitations alone, because the system-level impacts of teacher evaluations at the school level and beyond are such critical factors.

Although system-level evidence comparing, for example, districts placing high-emphasis on teacher observation vs. low-emphasis is not available, research has evaluated principals' use of observation data more broadly. As part of a study of principals' data use for human capital decision making in six districts, Goldring et al. (2015) report principals find "numerous productive uses [of observational tools] for decision making in their schools." Cannata et al. (2017), analyzing the six-district data along with Charter Management Organization administrators, report that 70% of principals utilized teacher observation data in making hiring decisions, while 60% reported teacher observations were "very important" in making hiring decisions. Yet, some principals question the reliability of observational data (Cannata et al., 2017, p. 210), and overall, studies find principals exercise much discretion in carrying out teacher observations as part of teacher evaluations, both in the number, duration, and formality of the evaluations and the extent to which the observations generate critical performance feedback (Cohen et al., 2019; Donaldson & Woulfin, 2018).

Even if teachers are primarily allocated to middle evaluation categories, and unreliably so, the observational frameworks themselves may still focus teacher attention and reflection on appropriate domains of instruction or enhance professionalization by providing a shared pedagogical language. For example, the principals in Cohen et al.'s study reported "… using the observation rubrics as ongoing frameworks for high-quality practice and useful tools for promoting more formative conversations about instructional improvement" (2019, p. 20). That is to say, unreliability does not appear to completely preclude positive impacts. Nor do certain forms of bias necessarily reduce the effectiveness of an evaluation system. Harris et al. (2013) show that school principals' valuing of teachers' sociability and organizational contributions to the school do affect principal ratings, but the ultimate impact of this "bias" on school effectiveness is not easy to predict since it involves factors school leaders understand to be important to school functioning.

Kelly (2012) argued that if teacher evaluation systems were to be implemented in such a way that large numbers of teachers were erroneously labeled as failures (as was the case with school accountability labels in the 2000s), this would be a policy disaster that would erode teacher

motivation and commitment. However, as shown in Table 1, across a wide array of protocols, most teachers end up as rated in the middle, and this appears to be the case in evaluative use as well. Kraft and Gilmour (2017) report that in many states less than 1% of teachers are rated as unsatisfactory, although that finding references overall ratings, not the observation component alone. Taking into consideration these low rates of negative evaluation, and existing evidence on use value, we cannot say that the major limitations outlined here preclude effective use in teacher evaluation.

Would incremental improvement in global protocols help address the limitations outlined here, further improving their use value? On this question we are more pessimistic. In the MET data we do find variation across protocols, suggesting some are more prone to specific limitations than others. Yet, all of the protocols suffer from each of these limitations to a greater or lesser extent, and we hypothesize that the overall quality of measurement may stem inherently from the ambitious, comprehensive goal of these protocols—to rate the entirety of a teachers' instructional effort. Thus, we conclude with consideration of an alternative to global protocols. We argue that the limitations of global observation protocols should no longer be accepted so readily, because fine-grained measures are emerging as an alternative. These tools reveal the limitations of global protocols in especially sharp relief. Moreover, unless a compelling alternative exists, many educational professionals and researchers may not be much persuaded to address the measurement limitations described here, incrementally or otherwise.

Fine-Grained Measures as an Alternative to Global Protocols

In studies of alcohol use, bold social scientists have pioneered the use of breathalyzers to collect fine-grained, occasion-specific measures of alcohol consumption as an alternative to traditional self-reported survey measures (Beirness et al., 2004; Smith et al., 2001; Wells et al., 1997).[12] Global observation protocols now give researchers and practitioners an occasion-specific view of classroom instruction, but they are not yet fine-grained. Fine-grained observational systems record and carefully analyze individual utterances, questions, turns at talk, etc., offering the potential for exceptionally reliable and precise teacher assessment and feedback. Fine-grained measures also offer the possibility of a more fundamental shift in the quality of information gleaned from classroom observation featuring: greater independence in assessing individual components of instruction; greater ability to identify teachers' own contribution to instruction; and an agnostic coding of instruction better suited to understanding teacher adaptation and change.

Historically, fine-grained measures of instruction, observational or otherwise, have been critical in documenting important basic features of American schooling, such as the low prevalence of genuine discussions in American classrooms (Nystrand & Gamoran, 1997), the wide variability in content and test standards in the US (Porter et al., 2011), or more recently, the content of texts read by diverse students (Northrop et al., 2019).

Yet, by their very nature, fine-grained measures tend to be difficult and expensive to collect/implement. Thus, in the past, due to their labor-intensive nature, such systems have been primarily used in research settings (e.g. Gamoran & Kelly, 2003; Howe et al., 2019; Murphy et al., 2009; Taylor et al., 2003) and in pre-service teacher preparation (e.g. Caughlan et al., 2013; Juzwik et al., 2013; Kucan, 2009).

---

[12] On November 3, 2019 the *New York Times* released a feature-length article about the unreliability of alcohol breath testing. Note however we reference the scientific use of breathalyzers in a study of a college norms campaign to reduce campus binge drinking and a case control study of recreational boating fatalities, as a fine-grained alternative to simple survey self-reports, not use in prosecuting individuals whose guilt should not be presumed.

Fine-grained systems also tend to narrow the focus of inquiry and inference. For example, while Nystrand and Gamoran's system of observation provided time summary statistics on more than a dozen basic classroom instructional formats (e.g. lecture, various forms of small group work, etc.), analyses of classroom discourse focused on only a few basic features such as authentic questions, uptake, and cognitive level. Indeed as Kachur and Prendergast's (1997) analysis in *Opening Dialogue* indicates, a class that seems non-dialogic based on Nystrand's primary indicators may in fact have an overall learning environment that takes student ideas seriously and generates high levels of engagement. In contrast, the global observational protocols in use today are exceptionally comprehensive. Some protocols even include elements of curricular planning that go beyond lesson observation (e.g. FFT). Moreover, beyond the protocol rubrics, documentation, and training materials, they elicit a qualitative, nuanced appraisal that additionally draws on the expert rater's internal frame of reference, memory, and training (Bell et al., 2014).

Fine-grained measures of instruction are thus promising but not fully tested. Automated methods under development by teams of educational researchers and computer scientists may soon overcome much of the inherent difficulty and expense associated with human observation and coding, which is a crucial step in making fine-grained measurement more widely available to researchers and practitioners (Kelly et al., 2018, Wang et al., 2014). For example, Kelly et al. (2018) demonstrated that it is possible to automatically detect and estimate the proportion of authentic questions in a class session with a reliability sufficient to complement or even replace human coding in research efforts. That result was obtained under technical requirements and constraints that preface wide spread use; only a teacher mic was used, without cameras or individually mic'ing students (see full discussion in D'Mello et al., 2015). More recently, these results have been replicated with teachers collecting data autonomously, without the need for research staff present (Stone et al., 2019). At the same time, experiments with a range of recording systems showed that the fidelity of the audio recording itself is important and must be evaluated in any automated measurement system; low or even medium quality audio may not yield sufficiently reliable estimates of classroom discourse properties.

Overall, the potential for widespread use of automated, fine-grained measures of instruction may mean that instead of incremental improvement of existing global protocols, researchers should pursue entirely new approaches. Yet, as promising as the automated systems sound, it is reasonable to wonder what might be sacrificed by a focus on more specific, discrete aspects of instruction? Even if fine-grained measures are inherently more precise and reliable, might they miss the forest for the trees, presenting a quantitatively accurate but qualitatively misleading portrait? Just as we have interrogated global protocols in this analysis, researchers must provide balanced evaluations of fine-grained measures of instruction that takes such possible limitations seriously and validate them on the many dimensions that affect robust use.

References

Aucejo, E. M., Coate, P., Fruehwirth, J., Kelly, S., & Mozenter, Z. (2018). Teacher effectiveness and classroom composition. CEPR Discussion Paper No. DP13166. CEPR.

Ball, D. L., & Cohen, D. K. (1999). Developing practices, developing practitioners: Toward a practice-based theory of professional development. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 30–32). Jossey-Bass.

Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, *59*, 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. In T. Kane, K. Kerr & R. Pianta (Eds.), *Designing Teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). Jossey-Bass. https://doi.org/10.1002/9781119210856.ch3

Beirness, D. J., Foss, R. D., & Vogel-Sprott, M. (2004). Drinking on campus: Self-reports and breath tests. *Journal of Studies on Alcohol*, *65*, 600–604. https://doi.org/10.15288/jsa.2004.65.600

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *The Elementary School Journal*, *111*, 7–34. https://doi.org/10.1086/653468

Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2007). Final reading outcomes of the national randomized field trial of Success for All. *American Educational Research Journal*, *44*, 701–731. https://doi.org/10.3102/0002831207306743

Brandt, C., Mathers, C., Oliva, M., Brown-Sims, M., & Hess, J. (2007). *Examining district guidance to schools on teacher evaluation policies in the Midwest region* (Issues & Answers Report, REL 2007-No. 030). U.S. Department of Education.

Campbell, S.L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal*, *55*, 1233–1267. https://doi.org/10.3102/0002831218776216

Camburn, E. M. (2010). Embedded teacher learning opportunities as a site for reflective practice: An exploratory study. *American Journal of Education*, *116*, 463–489. https://doi.org/10.1086/653624

Camburn, E. M., & Han, S. W. (2015). Infrastructure for teacher reflection and instructional change: An exploratory study. *Journal of Educational Change*, *16*, 511–533. https://doi.org/10.1007/s10833-015-9252-6

Cannata, M., Rubin, M., Goldring, E., Grissom, J. A., Neumerski, C. M., Drake, T. A., & Schuermann, P. (2017). Using teacher effectiveness data for information-rich hiring. *Educational Administration Quarterly*, *53*, 180–222. https://doi.org/10.1177/0013161X16681629

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

Cash, A. H., Hamre, B. K., Pianta, R. C., & Myers, S. S. (2012). Rater calibration when observational assessment occurs at larger scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly*, *27*, 529–542. https://doi.org/10.1016/j.ecresq.2011.12.006

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. G. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English*, *47,* 212–246.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, *18*, 947–967. https://doi.org/10.1016/S0742-051X(02)00053-7

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*, 378–387. https://doi.org/10.3102/0162373719893338

Cohen, J., Loeb, S., Miller, L. C., & Wyckoff, J. H. (2019). Policy implementation, principal agency, and strategic action: Teaching effectiveness in New York City middle schools. *Educational Evaluation and Policy Analysis*, *online first*, 1–27. https://doi.org/10.3102/0162373719893338

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.

Coburn, C.E., & Russell, J.L. (2008). District policy and teachers' social networks. *Educational Evaluation and Policy Analysis, 30*, 203–235. https://doi.org/10.3102/0162373708321829

Cor, M. K. (2011). *Investigating the reliability of classroom observation protocols: The case of PLATO*. Stanford University. Unpublished manuscript.

Croft, A., Coggshall, J. G., Dolan, M., & Powers, E. (2010). *Job-embedded professional development: What it is, who is responsible, and how to get it done well*. Issue Brief. National Comprehensive Center for Teacher Quality. http://files.eric.ed.gov/fulltext/ED520830.pdf

The Danielson Group. (2013). *The framework for teaching evaluation instrument*. Author. https://usny.nysed.gov/rttt/teachers-leaders/practicerubrics/Docs/danielson-teacher-rubric-2013-instructionally-focused.pdf

Darling-Hammond, L., & McLaughlin, M.W. (1995). Policies that support professional development in an era of reform. *Phi Delta Kappan, 76*, 597–604.

Darling-Hammond, L., Wei, R. C., & Johnson, C. M. (2009). Teacher preparation and teacher learning. In G. Sykes, B. Schneider, & D. L. Plank (Eds.), *Handbook of education policy research* (pp. 613–636). Routledge.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*, 81–112. https://doi.org/10.3102/01623737024002081

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*, 531–556. https://doi.org/10.3102/0162373718784205

D'Mello, S. K., Olney, A. M., Blanchard, N., Samei, B., Sun, X., Ward, B., & Kelly, S. (2015). Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015)* (pp. 557–566). ACM. https://doi.org/10.1145/2818346.2830602

Farkas, G., & Durham, R. E. (2007). The role of tutoring in standards-based reform. In A. Gamoran (Ed.), *Standards-based reform and the poverty gap: Lessons for No Child Left Behind* (pp. 201–228). Brookings Institution.

Gamoran, A. (2012). Improving teacher quality: Incentives are not enough. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 201–214). Teachers College Press.

Gamoran, A. (2013). *Educational inequality in the wake of No Child Left Behind*. Spencer Foundation Lecture to the Association for Public Policy and Management. APPAM. http://www.appam.org/awards/spencer-foundation-lecturship.

Gamoran, A., & Carbonaro, W. J. (2002). High school English: A national portrait. *The High School Journal*, *86*(2), 1–13. https://doi.org/10.1353/hsj.2002.0021

Gamoran, A., & Kelly, S. (2003). Tracking, instruction, and unequal literacy in secondary school English. In M. T. Hallinan, A. Gamoran, W. Kubitschek, & T. Loveless (Eds.), *Stability and change in American education: Structure, process, and outcomes* (pp. 109–126). Eliot Werner.

Gamoran, A., Nystrand, M., Berends, M., & Lepore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, *32*, 687–715. https://doi.org/10.3102/00028312032004687

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, *116*(6), 1–32.

Goe, L., Biggers, K., & Croft, A. (2012). *Linking teacher evaluation to professional development: Focusing on improving teaching and learning*. National Comprehensive Center for Teacher Quality.

Goldring, E., Grissom, J. A., Rubin, M., Neumerski, C. M., Cannata, M., Drake, T., & Schuermann, P. (2015). Make room value-added: Principals' human capital decisions and the emergence of teacher observation data. *Educational Researcher*, *44*, 96–104. https://doi.org/10.3102/0013189X15575031

Goldsmith, L. T., Doerr, H. M., & Lewis, C. C. (2014). Mathematics teachers' learning: A conceptual framework and synthesis of research. *Journal of Math Teacher Education*, *17*, 5–36.

Grossman, P. L., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J. H., Boyd, D. J., & Lankford, H. (2010). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. CALDER Working Paper No. 45. https://doi.org/10.1007/s10857-013-9245-4

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. https://doi.org/10.3102/0013189X14544542

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., Mashburn, A. J., Jones, S. M., Brown, J. L., Cappella, E., Atkins, M., Rivers, S. E., Brackett, M. A., & 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. https://doi.org/10.1086/669616

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal*, *51*, 73–112. https://doi.org/10.3102/0002831213517130

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. https://doi.org/10.1080/07370000802177235

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*, 794–831. https://doi.org/10.3102/0002831210387916

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*, 462–512.

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*, 253–263. https://doi.org/10.3102/0013189X14542154

Jackson, K., Rockoff, J. E., & Staiger, D.O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics*, *6*, 801–825. https://doi.org/10.1146/annurev-economics-080213-040845

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*, 199–218. https://doi.org/10.1086/376806

Joyce, J., Gitomer, D. H., & Iaconangelo, C. J. (2018). Classroom assignments as measures of teaching quality. *Learning and Instruction*, *54*, 48–61. https://doi.org/10.1016/j.learninstruc.2017.08.001

Juzwik, M. M., Borsheim-Black, C., Caughlan, S., & Heintz, A. (2013). *Inspiring dialogue: Talking to learn in the English classroom*. Teachers College Press.

Kachur, R., & Prendergast, C. (1997). A closer look at authentic interaction: Profiles of teacher-student talk in two classrooms. In M. Nystrand, *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (pp. 75–88). Teachers College Press.

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

Kelly, S. (2007). Classroom discourse and the distribution of student engagement. *Social Psychology of Education*, *10*, 331–352. https://doi.org/10.1007/s11218-007-9024-0

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 7–32). Teachers College Press.

Kelly, S., & Majerus, R. (2011). School-to-school variation in disciplined inquiry. *Urban Education, 46,* 1553–1583. https://doi.org/10.1177/0042085911413151

Kelly, S., Mozenter, Z., Aucejo, E., & Fruehwirth, J. (2019). *School-to-School Differences in Instructional Practice: Evidence from the Measures of Effective Teaching Study Data*. A presentation at the annual meeting of the American Educational Research Association, New York, April.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*, 451–464. https://doi.org/10.3102/0013189X18785613

Koedel, C., Mihaly, M., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *2015*, *47*, 180–195. https://doi.org/10.1016/j.econedurev.2015.01.006

Kucan, L. (2009). Engaging teachers in investigating their teaching as a linguistic enterprise: The case of comprehension instruction in the context of discussion. *Reading Psychology*, *30*, 51–87. https://doi.org/10.1080/02702710802274770

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*, 234–249. https://doi.org/10.3102/0013189X17718797

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *Elementary School Journal* 104(5): 409–26. https://doi.org/10.1086/499760

Liu, S., Bell, C. A., Jones, N. D., McCaffrey, D. F. (2019). Classroom observation systems in context: A case for the validation of observation systems. *Educational Assessment, Evaluation, and Accountability*, *31*, 61–95. https://doi.org/10.1007/s11092-018-09291-3

McCaffrey, D. F., Yuan, K., Savitsky, T. D., Lockwood, J. R., & Edelen, M. O. (2015). Uncovering multivariate structure in classroom observations in the presence of rater errors. *Educational Measurement: Issues and Practice*, *34*(2), 34–46. https://doi.org/10.1111/emip.12061

MET Project. (2018). User Guide to Measures of Effective Teaching Longitudinal Database. Bill and Melinda Gates Foundation.

Mihaly, K., McCaffrey, D. F., Staiger, D., & Lockwood, J. R. (2013). A composite estimator of effective teaching. (MET Project Research Paper). Bill & Melinda Gates Foundation.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessy, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740–764. https://doi.org/10.1037/a0015576

National Research Council. (2011). *Incentives and test-based accountability in education*. Committee on Incentives and Test-Based Accountability in Public Education, M. Hout and S.W. Elliott, (Eds.) Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. The National Academies Press.

New Jersey Department of Education. (2019). *Achieve NJ: Teacher evaluation and support*. Retrieved 12/03/2019 from https://www.nj.gov/education/AchieveNJ/teacher/

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic Pedagogy and Student Performance. *American Journal of Education*, *104*(4), 280–312. https://doi.org/10.1086/444136

Northrop, L., Borsheim-Black, C., & Kelly, S. (2019). Matching students to books: The cultural content of eighth grade literature assignments. *Elementary School Journal*, *120*, 243–271. https://doi.org/10.1086/705797

Nystrand, M., & Gamoran, A. (1997). The big picture: Language and learning in hundreds of English lessons. In M. Nystrand, *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (pp. 30–74). Teachers College Press. https://doi.org/10.2307/417942

Penuel, W. R., Riel, M., Krause, A., & Frank, K. A. (2009). Analyzing teachers' professional interactions in a school as social capital: A social network approach. *Teachers College Record, 11*, 124–163.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new U.S. intended curriculum. *Educational Researcher*, *40*, 103–116. https://doi.org/10.3102/0013189X11405038

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*, 4–15. https://doi.org/10.3102/0013189X029001004

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*(3), 523–553. https://doi.org/10.3102/00028312030003523

Reddy, L. A., & Dudek, C. M. (2014). Teacher progress monitoring of instructional and behavioral management practices: An evidence-based approach to improving classroom practices. *International Journal of School & Educational Psychology*, *2*, 71–84. https://doi.org/10.1080/21683603.2013.876951

Richards, J. K, Lounsbury, J. W., Skolits, G. J., Beavers, A. S., Huck, S. W., & Esquivel, S. L. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation*, *18*, 1–13.

Rothstein, J., & Mathis, W. J. (2013). *Review of two culminating reports from the MET project.* National Education Policy Center

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

Smith, G. S., Keyl, P. M., Hadley, J. A., Bartley, C. L., Foss, R. D., Tolbert, W. G., & McKnight, J. (2001). Drinking and recreational boating fatalities: A population-based case-control study, *Journal of the American Medical Association*, *286*, 2974–1980. https://doi.org/10.1001/jama.286.23.2974

Stacy, B., Guarino, C., & Wooldridge, J. (2018). Does the precision and stability of value-added estimates of teacher performance depend on the types of students they serve? *Economics of Education Review*, *64*, 50–74. https://doi.org/10.1016/j.econedurev.2018.04.001

Steinberg, M. P., & 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

Stone, C., Donnelly, P. J., Dale, M., Capello, S., Kelly, S., Godley, A., D'Mello, S. K. (2019). *Utterance-level modeling of indicators of engaging classroom discourse.* Proceedings of the 2019 International Conference on Educational Data Mining (EDM 2019): International Educational Data Mining Society.

Taylor, B. M., Pearson, D., Peterson, D. S., Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *The Elementary School Journal*, *104*, 3–28. https://doi.org/10.1086/499740

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education, 78*(1), 115–123. https://doi.org/10.1016/j.compedu.2014.05.010

Wells, J. K., Greene, M. A., Foss, R. D., Ferguson, S. A., & Williams, A. F. (1997). Drinking drivers missed at sobriety checkpoints. *Journal of Studies on Alcohol*, *58*, 513–517. https://doi.org/10.15288/jsa.1997.58.513

Wenzel, S., Nagaoka, J. K., Morris, L., Billings, S., & Fendt, C. (2002). *Documentation of the 1996–2002 Chicago annenberg research project strand on authentic intellectual demand exhibited in assignments and student work: A technical process manual.* Consortium on Chicago School Research.

White, M. C. (2018). Rater performance standards for classroom observation measures. *Educational Researcher*, *47*, 492–501. https://doi.org/10.3102/0013189X18785623

Yoon, K. S., Duncan, T., Lee, S. W-Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the evidence on how teacher professional development affects student achievement*. U.S. Department of Education.

Yuan, K., Le, V-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis*, *35*, 3–22. https://doi.org/10.3102/0162373712462625

About the Authors

Sean Kelly
University of Pittsburgh
spkelly@pitt.edu
Sean Kelly is a Professor in the Department of Administrative and Policy Studies at the University of Pittsburgh. He studies the social organization of schools, student engagement, and teacher effectiveness.

Robert Bringe
rbringe@live.unc.edu
Robert Bringe is a graduate student in the department of Economics as the University of North Carolina Chapel Hill. A former high school teacher, he is interested in studying policies related to teacher pay, teacher accountability and heterogeneous teacher effects

Esteban Aucejo
Arizona State University
Esteban.Aucejo@asu.edu
Esteban Aucejo is an Associate Professor in the W.P. Carey School of Business at Arizona State University, and a research associate at the Centre of Economic Performance. He studies economics of education, labor economics and human capital development.

Jane Cooley Fruehwirth
jcooleyf@live.unc.edu
Jane Cooley Fruehwirth is an Associate Professor of Economics at the University of North Carolina-Chapel Hill and a fellow of UNC's Carolina Population Center. She studies teacher effectiveness, social context, educational inequality and mental health.

About the Guest Editor

Audrey Amrein-Beardsley
Arizona State University
audrey.beardsley@asu.edu
Audrey Amrein-Beardsley, PhD., is a Professor in the Mary Lou Fulton Teachers College at Arizona State University. Her research focuses on the use of value-added models (VAMs) in and across states before and since the passage of the Every Student Succeeds Act (ESSA). More specifically, she is conducting validation studies on multiple system components, as well as serving as an expert witness in many legal cases surrounding the (mis)use of VAM-based output.

SPECIAL ISSUE
Policies and Practices of Promise
in Teacher Evaluation

education policy analysis archives

Volume 28 Number 62 April 13, 2020 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.

arquivos analíticos de políticas educativas
conselho editorial