education policy analysis archives

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

epaa aape
Arizona State University

Volume 31 Number 7 January 31, 2023 ISSN 1068-2341

Agnosticism in Instructional Observation Systems

Sean Kelly
University of Pittsburgh
United States

Citation: Kelly, S. (2023). Agnosticism in instructional observation systems. *Education Policy Analysis Archives, 31*(7). https://doi.org/10.14507/epaa.31.7493

Abstract: Many instructional observation systems are designed to provide rough, qualitative, highly-evaluative assessments on numerous core dimensions of teaching. Such systems achieve comprehensive overviews of teaching but are poorly suited to answering many discovery-oriented research questions. In contrast, fine-grained agnostic systems are needed to pose and answer causal questions about instruction, and to fully understand instructional variation and change. More speculatively, I argue that the agnostic quality of fine-grained systems may also be useful in promoting teacher learning. Agnostic systems offer choice, withhold judgement, make room for locally-compensatory practices, and promote a greater locus of control. Instructional observation systems that carefully and agnostically quantify instructional processes may best help teachers leverage their professional judgment and invigorate their professional practice.

Keywords: instructional practices; school/teacher effectiveness; observational research

Agnosticismo en los sistemas de observación instruccional

Resumen: Muchos sistemas de observación de la instrucción están diseñados para proporcionar evaluaciones aproximadas, cualitativas y altamente evaluativas sobre numerosas dimensiones fundamentales de la enseñanza. Dichos sistemas logran una visión general integral de la enseñanza, pero no son adecuados para responder muchas preguntas de investigación orientadas al descubrimiento. Por el contrario, se necesitan sistemas agnósticos de grano fino para plantear y responder preguntas causales sobre la instrucción y para comprender completamente la variación y el cambio en la instrucción. Más especulativamente, argumento que la calidad agnóstica de los sistemas de granularidad fina también puede ser útil para promover el aprendizaje de los docentes. Los sistemas agnósticos

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

Manuscript received: 3/1/2022
Revisions received: 6/21/2022
Accepted: 6/21/2022

ofrecen opciones, retienen el juicio, dan cabida a prácticas localmente compensatorias y promueven un mayor lugar de control. Los sistemas de observación de la instrucción que cuantifican cuidadosa y agnósticamente los procesos de instrucción pueden ayudar mejor a los maestros a aprovechar su juicio profesional y fortalecer su práctica profesional.
Palabras-clave: prácticas instruccionales; efectividad de la escuela/maestro; investigación observacional

Agnosticismo em sistemas de observação instrucional
Resumo: Muitos sistemas de observação instrucional são projetados para fornecer avaliações aproximadas, qualitativas e altamente avaliativas em várias dimensões centrais do ensino. Tais sistemas alcançam visões abrangentes de ensino, mas são pouco adequados para responder a muitas questões de pesquisa orientadas para a descoberta. Em contraste, sistemas agnósticos refinados são necessários para formular e responder questões causais sobre a instrução e para entender completamente a variação e a mudança instrucional. Mais especulativamente, argumento que a qualidade agnóstica de sistemas refinados também pode ser útil na promoção da aprendizagem do professor. Os sistemas agnósticos oferecem escolhas, retêm o julgamento, abrem espaço para práticas localmente compensatórias e promovem um maior locus de controle. Os sistemas de observação instrucional que quantificam cuidadosa e agnóstica os processos instrucionais podem ajudar os professores a alavancar seu julgamento profissional e revigorar sua prática profissional.
Palavras-chave: práticas instrucionais; eficácia da escola/professor; pesquisa observacional

Agnosticism in Instructional Observation Systems

In this essay I consider a fundamental property in the observational study of teaching: how agnostic is the measurement of instructional practice? Is instruction coded almost entirely along a continuum of effective practice? Or is salient variation in instructional practice measured more agnostically, such that the principal ways in which instruction varies are identified and carefully measured, without making assumptions about whether that instruction is effective or ineffective?

In practice, and to date, the agnostic (or the opposite, judgmental) quality of measurement tends to co-occur with the granularity of measurement (Hennessy et al., 2020); fine-grained coding systems focusing, for example, on the precise allocation of classroom time (time summary statistics) tend to be more agnostic than "global" observation protocols that provide rough, qualitative assessments on numerous core dimensions of teaching (Kelly et al., 2020). Indeed, the granularity of measurement may tend to generate agnosticism, where the specificity of coding allows the observer to set aside questions of impact or appropriateness. Such tendencies are not absolute, a fundamentally judgmental approach could still entail fine-grained measurement, and global codes could be agnostic (as in some of the codes of Cooley & Leinhardt, 1980 discussed below). More importantly here, I will argue that beyond the granularity of measurement, agnosticism itself is a core conceptual consideration in instructional observational systems. The balance of agnostic vs. judgmental approaches in research on teacher observation may have profound implications for future policy and practice.

Teacher Observation in Educational Policy and Research

Prompted by the federal Race to the Top initiative (RttT), in the 2010s 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 (New Jersey Department of Education, 2019). Observational protocols remain central to teacher evaluation policies, although the policy emphasis has shifted to formative feedback for instructional improvement as opposed to high-stakes, summative evaluation (Close et al., 2018). As this important change in *use* occurs, researchers will need to develop new observational tools to better match those goals.

Although difficult to document with certainty, it seems that the emphasis on using global observation protocols in teacher evaluation policy, and the research emphasis on developing and validating global observation protocols, were mutually-reinforcing. That is, interest among policymakers in summative evaluation motivated development by researchers of global observation protocols, while researcher claims of validity and reliability helped justify policy adoption of teacher evaluation with global protocols. This connection between policy and research is evident for example, in Kane and Staiger's (2012) *Gathering Feedback for Teaching* report from the Measures of Effective Teaching study, where the executive summary begins with a teacher evaluation motivation, and ends with implications, again, for teacher evaluation.

Yet, evaluation is just one use of observation protocols. As policy shifts from summative evaluation to more formative uses, another important development in teacher observation is occurring on the research side—the emerging possibility of automated methods of classroom observation (see e.g., Franklin et al., 2018; Jacobs et al., 2022; Jacoby et al., 2018; Jensen et al., 2021; Kelly et al., 2018; Liu & Cohen, 2021; McCoy et al., 2018; Ramakrishnan et al., 2021; Watson et al., 2021). For example, the TalkMoves system (Jacobs et al., 2022) records data with an iPad and a Swivl robotic camera base, along with five linked microphones arrayed around the classroom. This data is then automatically processed by TalkMoves, producing estimates of six talk moves (e.g., pressing for accuracy) as well as the overall ratio of student to teacher talk. As discussed by Ridge and Lavigne (2020) many of the policy challenges and possibilities for unintended negative consequences of teacher evaluation in the RttT era stemmed from the difficulty of administrators carrying out reliable, in-depth observations of teachers in person. Automated methods can in theory overcome those basic challenges, but first, a basic question must be addressed that will greatly affect the use of automated methods in policy and practice. What should the underlying coding scheme look like? More specifically, how fine-grained should it be, and how agnostic? Automated methods are developed in tandem with traditional human coding of instruction guided by protocols; indeed, it is the traditional human coding that provides the "gold-standard" referent and allows researchers to develop automated algorithms to analyze classroom audio and video data. Should the next generation of teacher observation protocols be judgmental or agnostic?

In some basic sense, "should-type" questions are related to intended uses (Goe et al., 2008), suggesting the need for varying observational designs to accommodate diverse uses and goals. In this essay I will also acknowledge several desirable properties that are relatively intrinsic to today's global observation protocols that promote important uses. In addition, while the granularity of measurement and the agnostic vs. judgmental approach are key axes of difference in the approaches I consider, it is also important to note basic similarities in approach. Throughout the essay, whether referring to agnostic or non-agnostic approaches, I am primarily concerned with instructional observation that uses existing codes, not exploratory research designed to identify new constructs and codes (See Derry et al., 2010 for discussion). That is, where observation is carried out using codes developed a-priori, not that emerge during coding.[1] Another important similarity is that both the agnostic and non-agnostic approaches described here can be used to study and document variation in opportunity to learn (Kelly et al., 2020b). In that way, both can be employed for discovery-oriented research to understand the sources of educational inequality. In contrast, in terms of uncovering the instructional practices that are most critical to learning and development, it is perhaps tautological that we should approach research more agnostically, focusing on instructional questions we don't already know the answers to. What may be less obvious is how an agnostic coding of instruction can inform teacher learning, an argument I develop in the final section.

In the remainder of the essay, I begin by considering in more detail what it means to conduct fine-grained, followed by, agnostic observations of teaching. I then discuss the relationship between the granularity of coding and agnosticism, which are definitionally distinct, but empirically correlated in existing systems of teacher observations, and I would argue causally linked, with small units of measurement promoting coding agnosticism. Next, I provide a brief (and highly selected) discussion of agnosticism in classic and contemporary classroom-observation research, providing concrete examples of agnostic approaches to teacher observation. Importantly, agnosticism is a matter of degree, rather than an absolute or binary distinction. I then conclude with the more speculative discussion of teacher learning.

What are Fine-Grained Measures of Instruction?

The term "fine-grained" is used to describe units of analysis/measurement that break the process and/or materials of classroom instruction into—in terms of the visual metaphor of grain-size—small pieces, as opposed to large chunks. Although this is a relative and qualitative rather than absolute and binary description, fine-grained measures of instruction would include measures at the level of individual seconds of time use, individual utterances (see discussion in Hennessy et al., 2020), or even various features within utterances, including acoustic features and individual words and phrases. Fine-grained measures also include primary units of analysis in texts and assignments used; for example, measures of text complexity based on exhaustive consideration of features of words, sentences, and the text as a whole.

Fine-grained measures may entail a simple binary coding[2] (e.g., was this second of class spent in small-group work, yes or no?), but result in a nearly continuous, ratio-scale measurement when used as summary statistics at the lesson level (e.g. 200 seconds of small-group is twice as much as 100 seconds, and an absolute zero value is well-defined). Fine-grained measures are often based on an *exhaustive* coding, as in the examples above. That is, every eligible unit is identified, and each and every unit coded for whether it meets criterion A, B, etc. I would also include measures produced through a *nomination* procedure that yield count data as fine-grained. For example, when a researcher reviews audio or video data and counts the number of times students are sanctioned, praised, etc. However, this does not produce fully satisfying ratio-scale data, because the denominator is

[1] There is an "inductive quality" to the idea of building up a portrait of instruction at the lesson- or teacher-level from many smaller, fine-grained observations. There is also an inductive quality to carrying out research on theoretically interesting instructional constructs, but ones where prior research offers competing perspectives on effects, where the effect can't be deduced ahead of time. Yet, the general framework I am discussing is deductive insofar as the codes are all known in advance, not developed during the study.

[2] Throughout this article I use the term "coding" to refer to the process of generating the values of an instructional variable from observational data. Similar to the process of survey response (Groves et al., 2009), this entails various cognitive processes, including judgement and estimation. Note that much of the literature relating to global observation protocols uses the term "rater" (see e.g., the MET Observation Measures Report), which has a much more evaluative connotation than "coder."

unknown or in a different scale. For example, using a nomination approach a researcher might identify three instances of praise in a 15-minute interval of instruction in each of two classes. Those classes would appear to offer identical levels of praise, but we don't know how many utterances were "eligible" for coding in those two different classes, and while there is a common denominator (15 minutes) it is not in the same units as the numerator and may ultimately make interpretation of the ratio somewhat ambiguous.

The term fine-grained does not refer to the number of categories in the response or rating scale applied to coding each unit, but again, to the units of analysis themselves.[3] For example, when coding the cognitive level of a question, task, etc., each unit might be scored simply low (simple reporting of information required) or high (analysis required), or instead, on a more finely differentiated response scale. While more finely differentiated response scales will in theory produce more precise measures, choice or construction of a response scale is a separate question from the units of analysis. Choice of a response scale should be based on the nature of variation in phenomenon under study, and how well the judgement process of coders can match that variation. For example, throughout Nystrand and Gamoran's program of research (discussed below), authentic questions were always conceptualized, as well as measured, as binary.[4] Other constructs, like "off-task behavior" are better conceptualized as continuously varying, based on the percentage of off-task students and just how distracted or disruptive they are, but there is a limit to a coder's ability to make those distinctions (see related survey methods research by Krosnick & Fabrigar, 1997). If binary codes for dozens or even hundreds of, for example, utterances, are aggregated to the lesson level, the resulting measure already provides quite precise distinctions (assuming the binary coding distinction does in fact capture meaningful differences in the construct itself). Overall, while the number of response categories may have some effect on measurement precision, response categories are a secondary concern compared to the more substantial distinctions in measurement considered here.

What is an Agnostic Approach?

In using the term "agnostic," I am describing instructional coding efforts that do not make assumptions about whether that instruction is effective or ineffective *at the point of coding*. Describing an approach to coding as agnostic has a similar meaning to describing an approach as "non-judgmental." However, I do not mean to imply that the information contained in the coding will never be used to make an inference about instructional effectiveness, or never be linked either in research or an applied setting to some outcome. Rather, the term agnostic is used in particular to highlight that the judgement does not occur at the point of coding even as it may occur in the future. For example, researchers might code teacher discourse for the occurrence of conjunctive or serial questions (Morine-Dershimer, 1985). It may be the case that in certain instructional contexts, increasing the prevalence of serial questions may lead to desirable outcomes. And with enough research, we may come to learn what those contexts are. However, an agnostic approach would simply code whether a given question was part of a series, without making any judgement about whether that series demonstrated excellence in teaching at that moment.

[3] Muijs et al. (2018) use the term fine-grained in their discussion of the International System for Teacher Observation and Feedback (ISTOF) to refer to a qualitative analysis structured by or informed by first scoring a lesson using ISTOF. In this case, the term does not correctly apply to the ISTOF itself as defined by the actual units of analysis (which is the entire lesson in the ISTOF).

[4] Arguably, there is still meaningful variation in "authenticity," and I have seen this arise in training raters to code for authentic questions (i.e., questions that seem partially authentic). Yet, our experience over time never suggested there was enough meaningful variation to justify revising the basic 0, 1 coding scale.

Importantly, while the judgmental quality of a protocol may be influenced by the anchoring/response labels, fundamentally, the judgmental quality pertains to or stems from the definition of the construct, not the labels. Consider the version of FFT used in the MET study, where each dimension of the protocol was scored on a 4-point scale labeled as unsatisfactory, to basic, to proficient, to distinguished. These labels are clearly very judgmental, and so, they contribute to that quality. Would the protocol become agnostic if those labels were changed to something like: low, somewhat low, somewhat high, high? Not in the case of FFT, because that protocol is defined as, and constructed with, the explicit goal of making *judgements* about teaching practice (Danielson, 2011, p. v). Thus, the constructs themselves are also defined, a priori, as being desirable and constituent components of teacher effectiveness. For example, there is no way that the domain "establishing a culture of learning" could in anyway be construed as agnostic, no matter the response/anchoring labels. That the outcomes are presumed in advance as part of the coder's judgement is made especially clear in FFT because the indicators and coding examples explicitly contain reference to student outcomes (see Kelly et al., 2020). Indeed, in some cases, it would not even by possible to change the anchor labels to "low" vs. "high" because that would change the meaning of the domain. For example, Domain 3a of FFT, "Communicating with students" is not about the amount of communication (as would be conveyed with a label like "high"), but rather characterized by judgmental adjectives like "rich" and "vivid." Overall, anchor/response labels will influence but not determine the agnostic quality of an observational system.

A second way in which the term "agnostic" seems to capture the approach I refer to in particular differently than a term like "non-judgmental," is in its relationship to interest or disinterest. Agnostic is not meant to mean "disinterested," and indeed, conveys an underlying interest, even as any ultimate judgement about the impacts of instruction at that moment are simply not present. In the case of serial questions for example, researchers may be specifically interested in the potential benefits of this discourse move in fostering widespread engagement, in checking for understanding, etc. But again, that interest does not factor into the coding; the coder simply has to decide whether the teacher question was part of a series or not. A third, very subtle connotation intended by use of the term agnostic, is that the approach I am referring to generally cannot be said to be immutably or perfectly non-judgmental. Indeed, I assume that in the real world of teacher observation, some judgement will always enter into the observational process. For example, if a coder, in the process of observation reaches the summary conclusion that what they are seeing from a given teacher is not high-quality instruction, and at the same time has made up their own mind that serial questions are generally beneficial, then this might cause them to underestimate the prevalence of serial questions. Likewise, the coder may believe that instructional Practice A tends to co-occur with Practice B; say for example, that serial questions tend to be of lower cognitive difficulty level than stand-alone questions. This too would introduce a judgmental quality to coding. However, these processes do not define the coding task as judgmental. Rather, they simply acknowledge that judgement may always contaminate a coding process.

Grain-size and Agnosticism

I developed my present preoccupation with agnosticism over the course of carrying out a program of research using fine-grained instructional observations, and at the same time as I was also part of a team of researchers working with the global observation protocols in the Measures of Effective Teaching Study data. Thus, I often contrasted these two research activities in my mind multi-dimensionally, as differing simultaneously in both grain-size and in agnosticism. Does such a multi-dimensional characterization of existing lines of research conflate two conceptually distinct features of teacher observation protocols? To be clear, these are, by definition, conceptually distinct features. The elaborated definition of agnosticism provided above does not include reference to grain size. Moreover, in the next section, I cite deviant case examples in the history of classroom observation, illustrating that agnosticism is not immutably coupled with grain size, in the past, and it need not be in the future. Indeed, even the coding in some of the global protocols used in MET were largely agnostic (see subsequent discussion of PLATO).

Yet, I maintain that the co-occurrence of fine-grained units of measurement and agnosticism in much research is no coincidence, that there is some causal link between the two that increases the likelihood that a fine-grained system of observation will be more highly agnostic. It is more than an accident that the running example above was the coding of individual serial questions (a quite fine-grained unit of measurement). What might be behind that causal linkage? There are three general possibilities. Agnosticism could promote selection of fine-grained units of analysis. The choice of fine-grained units of analysis could constrain or influence the generation of constructs that are coded. Or, both grain-size and agnosticism could be due to some other common cause; as would be true if they were both the logical result of a given research focus, that the researcher is focusing on discourse, or instructional time use, or other aspect of instruction.

All of these linkages seem possible, even likely to me, but here I will highlight the potential effect of grain-size on agnosticism. As the grain-size decreases, judgements of effectiveness would seem to become less and less reliable, and more and more contingent on the larger context of what that instruction is embedded in, in what came before, and what comes after. Even if the researcher is themselves convinced that a higher level of Construct A is on average, desirable, are they prepared to argue that each and every occurrence is by definition desirable, and that the coding task should be defined as making a judgement of effectiveness? In other words, once a researcher has decided on a fine-grained unit of analysis as the basis for coding, I believe that decision will tend to promote agnosticism in *what* is coded. Infusing judgements of effectiveness into fine-grained codes would too obviously reduce the validity and reliability of those codes.

As an example of the relationship between grain-size and appropriateness of judgmental coding, consider the use of PLATO Prime (containing 8 of the original 13 PLATO instructional elements) in the Measures of Effective Teaching study. Of the global protocols used in MET, I judge PLATO to be the most agnostic. Many of the codes pertain to the presence or frequency of specific forms of instruction and teacher moves, such as for example, the extent to which the teacher engages in modeling, or strategy use and instruction.[5] In this case, the system is not particularly fine-grained though, English/Language Arts instruction is scored over 15-minute instructional segments. As a result of early studies (Grossman et al., 2010), the PLATO developers realized that if this system were to be used in MET to produce inferences about overall teaching effectiveness, then the scores in each domain would only be valid (or much more valid) when aggregated across multiple segments over multiple days (MET Project, 2010), because lesson scores were highly sensitive to overall lesson context. For example, in the pre-MET pilot study (Grossman et al. 2010), with the exception of the modeling domain, scores were lower across the board in writing lessons than in literature lessons. Thus, even at the relatively large grain-size of 15-min intervals, it does not make sense to evaluate, for example, the presence or absence of strategy use as evidence of teaching effectiveness per se. It does however make sense simply to pose a question about how prevalent attention to strategy use was during that time interval (which is what the PLATO protocol does). In this example, an instructional domain (e.g., strategy use) appropriately maps onto a given unit of analysis (15 min intervals) when it is treated agnostically but not when it is

[5] Other codes in PLATO Prime, like the codes for behavior management and time management are more holistic and judgmental (i.e., is the behavior management effective).

judgmental.[6] I hypothesize that in general, finer-grained units of analysis require or encourage an agnostic approach to coding.

Other Features of Classroom Observation Protocols

Before proceeding with a continued discussion of agnosticism, I offer a brief discussion of other fundamental features of classroom observation protocols, summarized in Table 1. The six features in Table 1 are presented without reference to, or inclusion of traditional measurement concerns (reliability and validity). Table 1 implicitly reveals the limited focus of the present discussion; agnosticism is only one of many important features

Table 1

Some fundamental features of classroom observation protocols

Feature	Explanation/Definition	Example	Further Reading
Comprehensiveness	The extent to which the protocol encompasses a variety of separable (in theory) dimensions of teaching.	A protocol focused only on the nature of teacher talk lacks comprehensiveness. A protocol focusing only on evidence of dialogism in teacher talk is even less comprehensive.	Praetorius & Charalambous (2018) provide a content analysis of protocols and the dimensions and constructs they measure
Grain-size of units of measurement	The extent to which a protocol identifies and utilizes the finest (smallest) units of measurement at which a construct can be robustly scored.	A protocol classifying activity structure in 15-min increments is courser-grained than a protocol classifying activity structure in 1-second increments.	Hennessy et al. (2020) discuss grain-size considerations in the coding of classroom dialogue.
Agnosticism	The extent to which codes on a given variable are defined without reference to, or presumption of effectiveness, such that no judgement of effectiveness, appropriateness, etc. is made *at the moment of coding.*	A protocol classifying a given domain as "positive" (e.g., "positive behavioral climate") is inherently judgmental and lacks agnosticism. A protocol counting the number of times a teacher references classroom rules is agnostic.	See present article

[6] Technically in this example the coding itself does not vary (the authors are commenting that judgmental uses become more feasible as smaller units of analysis are aggregated to larger units), but I infer that if a judgmental use would be inappropriate, then so would building that overall judgement of effectiveness into the coding itself.

Feature	Explanation/Definition	Example	Further Reading
Formative (vs. reflective) emphasis	The extent to which a protocol presumes internal consistency. Relatedly, the extent to which dropping a measure changes the meaning of the overall scale score	Protocols conceptualized as formative may have identical surface features to protocols conceptualized as reflective. Yet, a formative conceptualization emphasizes that no internal consistency is assumed, but instead, the set of measures collectively constitute, rather than reflect, [effective, if judgmental] teaching practice.	Jarvis et al. (2003) provide an overview of this distinction. White et al. (2021) discuss this at length with regard to teacher observation. See also Kelly et al. (2020).
Preoccupation with a specific locus of authority	The extent to which a protocol presumes a student-centered (developmental) perspective on authority as opposed to a teacher-centered (incorporative) perspective on authority, or vice-versa, is either fundamentally related to effectiveness (if judgmental) or otherwise fundamentally of interest (if agnostic but still focused on expressions of instruction that reflect a developmental perspective on authority).	Protocols that assume a particular locus of authority is desirable are imbalanced in focusing only on the positive expression of that authority. For example, protocols heavily informed by theories of dialogism may assume a developmental perspective on authority is desirable, and contain only measures related to the positive expression of that perspective.	See Pace & Hemmings (2007) and Metz (1978) for general discussion, Kelly (2010) for relation to instructional measures, and Lehesvouri et al. (2018) for discussion of observational research.
Focus on teacher moves (vs. realized nature of instruction)	The extent to which a protocol more narrowly emphasizes what teachers do, or "teacher moves," as opposed to the realized nature of instruction	A protocol measuring the ratio of student to teacher talk has measured the realized nature of instruction, where the observed ratio is due to a blend of teacher and student influences.	See Kelly et al. (2020) for discussion of this feature in global protocols. See Russ et al.(2016) for discussion of teaching conceptualized as a set of actions.

Beyond agnosticism and grain-size, the comprehensiveness of a protocol, along with the formative (vs. reflective) emphasis, the extent to which the protocol is preoccupied with a specific locus of authority, and the relative emphasis on teacher moves (vs. the realized nature of instruction) are all key features of classroom observation protocols. Literature discussing these features is referenced in Table 1. These features are sometimes discussed in documentation accompanying the development of a protocol (e.g., a grounding in student-centered theories of instruction is often cited), and in other times are notably absent, as in the case of whether protocols have a formative or reflective emphasis. Subsequently, in my view, a great deal of research then focuses narrowly on traditional measurement properties (i.e., the most common expressions of reliability and validity), without much consideration of the basic features outlined in Table 1 and how they affect the results. However, this focus is beginning to change, as researchers interrogate how these features might function (Hennessy et al., 2020; White et al., 2021). The present discussion seeks to add to that changing focus by highlighting agnosticism in particular.

A Look Back at a More Agnostic Era of Measurement: 1970s Era Process-Product Research

A remarkable amount of process-oriented, quantitative research on classroom instruction occurred in the 1970s and surrounding years (Brophy, 1986; Dunkin & Biddle, 1974; Rosenshine, 1971). In this research, the unit of analysis was often well-defined and exhaustive (e.g., every second of instruction was classified as X or Y), or alternately, the coder nominated events/instances as having occurred, which produced counts of events over an interval of time, even as the unit of analysis, or the number of units evaluated, was less clear. Another feature of this research is that it often involved a very close correspondence between measures of achievement and instruction, which was useful for demonstrating important variation in opportunity to learn in schools. For example, Barr and Dreeben (1983) documented the number of vocabulary words students were exposed to, and hence learned, in ability-grouped classrooms (see e.g., Arehart, 1979; Good et al., 1978 for other examples of content coverage). In contrast, modern studies of instruction and learning using global instructional protocols focus on much more comprehensive dimensions of teaching, and much more comprehensive, standardized test outcomes (Aucejo et al., 2022; Mihaly et al., 2013).

A third feature of 1970s era process-product research is that it often (but not always) incorporated fine-grained, agnostic measures of instruction. Some major process-product studies used a mix of global (including teacher reports) and fine-grained (based on coding of video-tapes) approaches (Cooley & Leinhardt, 1980). But much of the process-product research focused on a careful coding of instructional time use, or counts of particular instructional events, without building inferences about the desirability of that time use or events into the codes themselves. For example, Smith (1979) conducted a fine-grained coding of teacher questioning, the percentage of content-related teacher questions in an Algebra unit, that was carried out directly from audio tapes, without transcription, to aid coding efficacy. In another example, Smith and Land (1981) summarize the results of 12 fine-grained studies by their team and others where the vagueness of teacher speech was quantified as a rate/frequency of instances per minute.

In these cases, although researchers might have hypothesized in advance that vagueness would reduce student learning, or a greater incidence of content-related questioning would enhance learning, these codes were agnostic *at the point of coding*. That is, the coder was not making a judgement of effectiveness at the time of coding. Indeed, vagueness included sub-categories of terms including probability-related terms (e.g., “frequently,” “in general,” etc.), which one might view as desirable or even essential to cultivating statistical thinking (Ritchey, 1999). Importantly, this agnostic quality overlaps with but is not the same as the distinction between low- vs. high-inference coding of teacher behaviors (Rosenshine, 1970, 1971). Rather, specifically, it refers to a low, or ideally completely absent, inference about *linkages with outcomes in particular*. As the next section shows, there are many examples of fine-grained, agnostic coding that require a substantial amount of judgement by coders and are simply quite difficult to code, and thus are not low-inference in the more holistic sense.

The agnostic quality of much of the 1970s era process-product research was I believe, instrumental to reaching nuanced, non-obvious understandings of instruction. Consider for example, Brophy's (1986) summary of findings on the cognitive level of questions posed to students. Again, while researchers might have a priori expectations about such effects, the actual coding of cognitive level for an individual question is an evaluation orthogonal to the ultimate effect of that question (but at the same time this is by no means an easy construct to code). Brophy argues that not only did the research on the cognitive level of questions not reveal that more high cognitive level questions is uniformly better, if anything it reinforced the value of less demanding questions.[7] Overall, according to Brophy, process-product research validated several abstracted principles of instruction, such as the importance of active instruction, but also showed how complicated, and goal and context dependent teaching is.

Nystrand and Gamoran's Computer-Based Coding Platform

The influence of process-product research was evident in the subsequent generation of instructional research, including Nystrand and Gamoran's computer-based coding platform (called CLASS, but to avoid confusion with other research I will refer to it as the N&G system). The N&G approach is an important example of an instructional observation system not only because it is well known and influential, but because it is exceptionally fine-grained, and fundamentally agnostic even in spite of its developers' belief in the value of dialogic instruction. This system was used in English language arts and social studies classrooms in a series of studies beginning in the late 1980s and culminating in the Partnership for Literacy study in the early 2000s (Gamoran & Kelly, 2003; Gamoran & Nystrand, 1992; Juzwik et al., 2008; Kelly, 2008; Nystrand & Gamoran, 1997). The N&G system had slight variations over time (e.g., changing codes for varieties of uptake), but the basic system can be described as follows. First, the coding platform was computerized, with an interface to make coding efficient and minimize errors and omissions. Second, instructional time use was exhaustively coded into 14 mutually exclusive activity segments: procedures and directions, discipline, classroom interruptions, lecture, question and answer, discussion, reading aloud, silent reading, role play, games, tests and quizzes, student presentation, seatwork (further sub-divided according to teacher involvement), or small group work, with anything else coded as other. Third, in addition to these codes, which produced time summary statistics (e.g., number of seconds spent in discussion), certain properties of questions were coded, including source (teacher or student), response (was the question answered), nature of evaluation in the response, authenticity, uptake, and cognitive level, thereby producing question property statistics (e.g., the proportion of teacher questions that were authentic, etc.).

A number of basic properties are evident in this early and influential computerized coding platform. First, while the coding system is fairly well-elaborated, it is hardly comprehensive of dimensions of instruction that might be related to learning outcomes. For example, teacher affect/emotion is entirely omitted (apart from any inherent association with evaluation, uptake, etc.), as are many cognitive processes related to development in literacy and other subjects that are well-developed in other programs of research (e.g., elements of argumentation; Connor et al., 2014; Reznitskaya et al., 2001; Wilkinson et al., 2010; or differentiated instruction, see e.g., Van Geel et al.

[7] Findings on cognitive level paralleled findings on question *difficulty*, a related but conceptually distinct concern with the rate at which students are able to offer a correct response.

2019). Indeed, the focus of statistical analyses with data generated from the system was often further narrowed to the sub-set of measures most closely related to dialogic instruction (discussion, question authenticity, and uptake). The most influential finding from this work is perhaps a very simple one; the low incidence of genuine discussion in secondary ELA classrooms (Gamoran & Nystrand, 1997).[8] Second, the N&G coding scheme is exceptionally fine-grained, with time exhaustively coded in seconds, and every single instructional question during Q&A segments recorded, averaging about 70 questions per class session (Juzwik et al., 2008). As a result, and even with the decision to not code *all* discourse, entering and revising codes took more than three hours per class session.

Third, this coding system was a mix of low and high-inference codes, with difficult subjective judgements involved in the distinction between Q&A and discussion, question authenticity, uptake, and cognitive level among others. Fourth, while the program of research was almost always closely linked with dialogic theories of instruction, and explicit beliefs in the value of dialogism among the research team, the coding itself was agnostic with respect to outcomes.[9] That is, in coding a question as incorporating uptake or not, the coder was making no judgement about the timeliness, necessity, or appropriateness of that uptake. Indeed, this agnosticism is clearly present in the findings of this program of research. Indicators of dialogism are not uniformly, consistently, or strongly linked with achievement growth, and positive effects were often context dependent.[10]

Global Protocols for Teacher Accountability

Following efforts by researchers like N&G who carried out large-scale research with agnostic, fine-grained observational systems, a new generation of observational protocols was being developed that provided much more global and valenced assessments of instruction. Global protocols include cross-subject systems such as the Framework for Teaching (FFT), Classroom Assessment Scoring System (CLASS), Teaching for Robust Understanding of mathematics (TRU), as well as subject specific ones such as the Mathematical Quality of Instruction system (The Danielson Group, 2011; Hamre et al., 2013; Hill et al., 2008; Schoenfeld, 2014).

It is difficult to find statements of intended use associated with the development of these protocols. To preface this, I don't recall N&G ever providing a discussion of appropriate use of their system; while it was clearly designed for discovery-oriented research purposes, it also found uses in teacher education (Caughlan et al., 2013) and could conceivably be used in other applied ways. Likewise, I have rarely found "use this observational system for X but not Y" type statements in articles, reports, or manuals accompanying the *initial* development of global protocols. One exception occurs in the Protocol for Language Arts Teaching Observation (PLATO) tool literature, where the developers describe it as useful for (1) research on teaching, (2) teacher development, and (3) clinical supervision (Grossman et al., 2010, p. 26). Another exception is the Classroom Assessment of Sociocultural Interactions (CASI), whose developers describe as promising in exploratory research, in the development of interventions, in teacher preparation, in professional development, but not in summative evaluation (Jensen et al., 2018). Likewise, in a later analysis, Bell

[8] Not to undersell the varied implications of this research, which included ground-breaking inferences about how teacher discourse moves build toward dialogism (Nystrand et al., 2003), how instruction differs across tracked classrooms (Gamoran et al., 1995; Gamoran & Kelly, 2003), and fundamental inferences about the distribution of student engagement (Kelly, 2008).

[9] For some evidence on researcher perspectives in this line of research see for example the well-developed discussions in the first Chapter in Nystrand's *Opening Dialogue* on why/how dialogism promotes learning, or Gamoran & Nystrand (1992).

[10] As a good example of this, consider the findings reported in Gamoran et al. (1995).

et al. (2012) caution against summative use of the CLASS-S system.[11] My own view is that in developing a research tool, it would be surprising if the developers articulated a narrowly tailored utility for the tool. If the tool provides any kind of useful insight into teaching, then it might be useful for many different purposes, and scientifically, why preclude investigation of possible uses.

If the espoused use of global observational protocols is unclear or multipurposed, they were quickly put to one particular use, teacher accountability. One interpretation of the development and use of global protocols for teacher accountability is that these tools helped translate the insights of a huge body of educational research into tools for instructional improvement. This includes the basic insights of the process-product era of research, and definitely includes insights from Nystrand and Gamoran's research (domains in global protocols related to teacher questioning are highly consistent with that research). For myself, I find many of the global protocols offer an impressively clear and helpful organization of instructional processes.[12] They are also quite comprehensive, even as some protocols have special emphases. When used for evaluation, it is useful to have a comprehensive observational protocol, to reduce construct under-representation (i.e., important domains of practice not captured in the protocol). Praetorius and Charalambous (2018), comparing 12 observational systems in use in the US and Europe, find that none of the frameworks are comprehensive of all constructs used in others, and that for similar constructs they differ somewhat in the elements that comprise key instructional constructs. Likewise, Jensen et al. (2019) argue the protocols used in MET were in fact not comprehensive of the full range of teaching practices affecting achievement and were not necessarily theoretically/conceptually coherent. Nevertheless, many educational researchers probably view this set of tools as a logical, timely, and generally coherent and well-developed extension to more discovery-oriented approaches to classroom observation.

How global observational protocols function in different approaches to school improvement depends in part on their measurement properties. In various ways, if they are flawed measures, this may negatively impact use. There is a literature investigating the measurement properties of global observation protocols (Bell et al., 2014; Campbell & Ronfeldt, 2018; Cohen & Goldhaber, 2016; Gitomer et al., 2014; Humphry & Heldsinger, 2014; Liu et al., 2019; McCaffrey et al., 2015; White, 2018; White et al., 2021), and some of that literature offers very sharp critique (Kelly et al., 2020). To describe just one concern, scores on the various sub-domains of global protocols, ostensibly important to providing teachers useful, domain-specific feedback, are not anywhere near as separable in practice as the structure of the protocols would make them out to be (Aucejo et al., 2022; Humphry & Heldsinger, 2014; Liu et al., 2019; McCaffrey et al., 2015). Yet, even an imperfect measure of classroom instruction can be very useful and have uses beyond the scores themselves. For example, Goldring et al. (2015) find that apart from the utility of actual scores in evaluation or staffing (which may be unreliable and clustered in the middle of the scale), principals report that the

[11] Using a decomposition of variance analysis with 4-5 CLASS observations per teacher, Bell et al. (2012) caution against making generalized inferences about the average level of teaching for the selected classroom for the school year, and relatedly, various high stakes decisions that might accompany such generalizations including merit pay, certification, etc. However, I did not find the consideration of the lesson-level variance in particular as problematic for this form of generalization (although other sources were, so Bell et al.'s overall point may still hold). I interpret lesson level variance as a reliability concern at the teacher-level, not a validity concern (i.e., lesson to lesson variance is expected, and as long as a sufficient number of observations is conducted, variance at the lesson level will average out and the unreliability will be tolerable).

[12] Along with that impression, I also wonder about the level of abstraction in a given protocol. Why four domains, why five domains? Why not just two, as in for example, Shernoff's (2013) theory of environmental complexity? What is most useful? A given protocol has to settle on a set level of abstraction for the main scoring domains, even as different levels may be differently useful for different purposes or different stages of teacher learning.

observational frameworks themselves do focus teacher attention and reflection on appropriate domains of instruction and enhance professionalization by providing a shared pedagogical language. Thus, the overall use of the observational tool is not restricted merely to producing scores.

One of the most important shifts in the global protocol era was from fine-grained measurement of individual seconds of time, individual questions, even individual words, to a much more course-grained, global qualitative coding. Such course-grained measurement does not lend itself to promoting incremental, continuous improvement. Additionally, score distributions are not often very uniform (in the statistical sense of the uniform distribution). In such systems, most teachers, most of the time are just scored as "in the middle" (Kelly et al., 2020; Kraft & Gilmour, 2017). Global protocols are unlikely to detect small improvements or other changes in teaching. Yet, my central point in this essay is that an even more important shift, *the most important shift*, was the shift from an agnostic coding to a valenced, judgmental coding. The judgmental quality of global observation protocols is entirely consistent and fitting with the use of those protocols in teacher evaluation. At the same time, it limits these tools' ability to continue to build causal understandings of teaching processes. If the coder, by definition, must make an assessment about what is effective at the time of observation, the utility of the protocol is limited primarily to describing variation in effectiveness, rather than adjudicating competing perspectives/hypotheses on what practices are effective.[13] Additionally, I will argue below that agnosticism is a valuable property in promoting teacher learning. Agnosticism is of no use in evaluation for accountability purposes, but what may be overlooked, is that it can be useful in feedback and instructional improvement. To preface this argument, I first consider recent technological advances that make feedback with instructional observation systems more efficient.

Contemporary Fine-Grained Systems in the Automated Era

One trait that classroom observation systems/tools have shared, from the early process-product research, to research in the 1980s and 1990s, to the global protocols of today, is that they are just fundamentally labor intensive and costly. A trained observer/coder has to go to a class, or sit down and watch a video, and make judgements that are challenging in their complexity or just in the concentration required to review and code a lot of content. Attention has been paid to streamlining this process, and there has been almost continuous deliberation in all of these literatures about how many teaching constructs to include in a coding system, how much of a lesson to code, and when making inferences about teachers' stable instruction, how many lessons to code per teacher (Van der Lans et al., 2016). But even if the coding itself is streamlined, and aided by computerization, it is still labor intensive and subject to the risks of coder fatigue, bias, or even mischievousness.

Automated systems of observations, based on audio and/or video data, where teachers themselves autonomously collect and submit data online for processing, offer a truly remarkable increase in efficiency of classroom observation. In the audio-based research my colleagues and I are carrying out, when the recording equipment and software expenses are amortized over an intensive, long-term program of research in schools, districts, or beyond, the cost is orders of magnitude cheaper than using university or school-system based data collectors.

Measurement challenges in classroom research, whether human or automated, go far beyond simple questions about reliability. For example, does the system focus on the teacher's contribution to instruction, the enacted quality of instruction students experience, or some unknown combination thereof? Considering such grand challenges, it remains to be seen how well automated systems

[13] Global protocols can also in theory investigate differential importance of effectiveness across sub-domains, or context-dependent effectiveness (e.g., interactions between classroom composition or other contextual variables and a given sub-domain).

function compared to human coded systems, although there is at least now evidence some important instructional constructs can be reliably coded (Kelly et al., 2018; Jensen et al., 2020; Jensen et al., 2021). Yet, beyond/before these basic challenges, a huge question looms in the development of automated systems of classroom observation; will these systems be based on global coding protocols, or will a more fine-grained, agnostic approach provide the gold-standard human codes that automated systems use to classify classroom instruction? While it is of course possible to eventually pursue both of these methods/goals, given limited resources for educational research, I encourage a focus on fine-grained, agnostic approaches.

Four Principles of Agnostic Classroom Observation for Research and Teacher Learning

No research is ever fully agnostic. Even noticing salient variation in the first place raises the question of why it is deemed to be salient. Nystrand and Gamoran were coding authentic questions because of an interest in dialogism. But classroom observation research of the process-product era and beyond was agnostic *enough* to shed light on competing perspectives on classroom instruction. What competing perspectives can such research address? Does whole-class instruction, as opposed to small group and individualized instruction (activity structure) really create a classroom motivational climate risky for lower-achieving students? Does the cultural content of instruction really affect student engagement? Is too much teacher-centered instruction detrimental to learning or school belonging, and if so, how much is too much? Agnostic, balanced, fine-grained research is the only way to generate real answers to these questions, even as individual studies offer only incomplete/partial views into these questions.[14] Researchers can't answer any of these questions using a global observation protocol because the judgement of effectiveness is made at the time of coding.

What may be less obvious is how agnosticism in a situation where teachers receive feedback on instruction might be beneficial to teacher learning. Overall, I find Clarke and Hollingsworth's (2002) interconnected model of teacher professional growth useful in considering feedback with teacher observation systems. In this model, the teacher observation system serves as an external source of information and stimulus, which supports reflection and enactment, including "professional experimentation." The context in which observational systems provide external information and stimulus could range from pre-service teacher training, to individualized "fit-bit" style use by teachers, to professional study groups, to whole-school interventions. Beyond this basic model, I do not view observational systems as necessarily restricted to a specific conceptualization of what teachers should be reflecting on or on what is most foundational to the teaching process over the long-run (i.e., teaching as a set of actions, teaching as a way of thinking, teaching as interacting, etc., see discussion of these in Russ et al., 2016). Observational systems can provide external information that might function in many different paradigms.

Both global protocols and more agnostic fine-grained protocols provide teachers with that critical external source of information missing in day-to-day practice (Clarke & Hollingsworth, 2002; Goe et al., 2012) and both can help teachers identify weak points in their instruction. Both offer a focusing-effect on key instructional constructs, although global protocols, due to their comprehensiveness and abstraction, do so in a different way. Fine-grained protocols, due to precision in measurement alone, may be a better platform for encouraging teacher experimentation (in the Clarke & Hollingsworth sense). Apart from these features which are more or less shared,

[14] "Balance" is a more generic property of research that goes beyond issues of classroom observation. By balance I mean that the research is positioned to show both the positive and negative effects of an educational policy or practice. I often point to the Hamilton et al. (2007) implementing standards based accountability report as a particularly balanced study, although that uses a survey methodology.

agnosticism may be useful in promoting learning by increasing teachers' receptivity to receiving feedback through four mechanisms.

Choice

First, an agnostic system of measurement presents teacher users with *choice*. Global observation protocols assume that each sub-domain contributes to effective teaching, and so by definition, it would be wrong not to focus on *all* of the domains scored on the protocol. In contrast, an agnostic system presents teachers with information on the nature of their instruction, five constructs, 10 constructs, or however many are measured, without insisting that every single one of those constructs maps onto effectiveness in an obvious way in every single lesson. Instead, it offers teachers choice, where teachers can focus on what they themselves want to focus on. That focus can change over time. Choice is limited by what the system measures, but nevertheless is non-trivial. If choice is a basic determinant of engagement (Assor, 2012; Cordova & Lepper, 1996; Ryan & Deci, 2000), then this is an important property of a feedback system.

The presence of choice in agnostic observation systems was recently highlighted in two small scale user-studies of the TalkMoves application (Jacobs et al., 2022) and the Teacher Talk Tool (Jensen et al., 2020). Jacobs et al. (2022) report that mathematics teacher users were much more attentive to one system output, the ratio of student to teacher talk (which is also a key output reported in the TeachFX system), than to other talk moves such as revoicing, pressing for accuracy, etc. Somewhat differently, Kelly et al. (in progress) found that in user-study interviews, English teacher users demonstrated substantial person-to-person variability in attention to specific discourse features reported by the Teacher Talk Tool. For example, some users were highly attuned to feedback on authentic questions, while other users did not focus at all on that feedback. These results demonstrate that when given the opportunity, teachers appear to exercise choice in the particular dimensions of instruction they themselves choose to analyze and act on compared to what information the system offers. At least in these current systems however, these choices are highly constrained by the few instructional constructs the system provides feedback on. These systems do not yet give teachers the agency to pose their own questions about instruction.

Withholding of Judgement

Second, and even more intrinsically, agnostic systems *withhold judgement*. This withholding is essential, but also limited in two major ways. First, there is some judgement inherent in quantifying a teaching construct as "low" or "high," and certainly some judgement/comparison if the system provides a percentile ranking. Some teachers, at the end of the day, will *feel* they are judgmental. Second, the system, or rather, the researchers who developed it, have decided what is important to count (see limitations of choice above). Yet, there is some element of withholding judgement inherent to systems that define measured constructs in a value-free way. For example, when a teacher is told they scored low in offering students praise, or that they scored very low, in the bottom 5%, of offering praise, this is not the same thing as telling the teacher they *should* have offered more praise, or that, by definition, the lesson was ineffective because they offered less praise.

Withholding judgement may be a basic principle of persuasion/attitude change. I am not aware of performance evaluation studies that address this question. However, classic theories of attitude change warn that attitudes often serve an ego-defensive function (Katz, 1960) and that an individual is more likely to be receptive to an argument (e.g., that one might need to improve in an area of teaching) if it is closer to their existing attitudes and beliefs (Sherif, 1963). For example, even if a teacher really should offer more praise, they may be more likely to come to that conclusion if left to deliberate on their own, rather than by adopting a more didactic approach. The principle of withholding judgement is used in social norms campaigns, which have been shown to be effective in changing risky behavior (Foss et al., 2003; Haines, 1996). While this may seem like splitting-hairs, in persuasion and attitude change, it is worth considering that a light touch may be preferred to a heavy-handed approach.

Locally-Compensatory Practices

Third, and related to the construct of choice, agnostic systems accommodate *locally-compensatory* teaching practices. Recall the N&G finding that very little genuine discussion occurs in secondary ELA classrooms (less than 2 minutes in the earliest research, and a mere 45 seconds in the control classrooms in the Partnership for Literacy Study, Juzwik et al., 2008). Reading researchers are sometimes skeptical of the emphasis placed on discussion in dialogic instruction, arguing that active, direct time spent reading is the most important determinant of literacy development, especially among struggling readers.[15] But the N&G research shows that the amount of discussion could be doubled, tripled, or more, and still leave plenty of time for the vast majority of instruction to be more direct. In a similar way, agnostic systems leave room for teachers to engage in locally-compensatory practices that are atypical of local instruction, and offer students a different learning modality or experience. In other words, they allow for a "students at my school don't get a lot of xxxx; I offer that to them in *my* class" type approach. Agnostic systems don't say that any given teacher must offer a certain instructional approach.

Internal Locus of Control

Fourth, agnostic systems, especially automated ones, may offer teachers a greater (more internal) *locus of control* concerning system outputs and future change or improvement in measured teaching constructs. Locus of control is an efficacy-related concept differentiating beliefs/perceptions about control over outcomes (Furnham & Steele, 1993; Rotter, 1966). Individuals with an internal locus of control believe outcomes are strongly determined by their own actions as opposed to external factors like luck or the decisions of others (Schunk, 1991). Locus of control has long been theorized as a central component of teacher efficacy and related constructs like burnout (Dworkin, 2009; Zee & Koomen, 2016). When an external observer scores instruction according to a researcher-developed criterion, it may be easy for the target of that observation (the teacher) to view the resulting scores as primarily determined by the scorer's own preoccupations, biases, etc., rather than anything they themselves might do. When that observer is a local actor, locus of control may be especially affected by pre-existing trust or mistrust. Stated differently, agnostic systems are less likely to be associated with system alienation attitudes and beliefs than evaluative systems, particularly when that evaluation is carried out by a local administrator. Teachers are highly educated, and education is the most important predictor of (reducing) misanthropy (Smith, 1997). But teachers are still vulnerable to feelings of general mistrust, and to mistrust in the occupational system they are embedded in and to the administrators they work with (Price, 2021, 2012; Tschannen-Moran & Hoy, 1998; Van Maele & Van Houtte, 2009). A non-trivial number of teachers may believe that the administrators who review their work "just don't like anything I do."[16]

[15] This argument has been made to me, anonymously, by IES reading panel reviewers.

[16] The literature on faculty trust in principals does not hone in on the prevalence rate of gross mistrust of principals. There is a developing literature on teacher perceptions of administrators' classroom observations in particular (Cherasaro, et al, 2016; Grissom et al., 2018; Kraft & Christian, 2019; Pepper et al., 2015) as well as administrators' own experience (Donaldson & Woulfin, 2018; Hunter & Rodriguez, 2021; Jones et al., 2021; Kraft & Gilmour, 2016). In the former, there is a wide distribution of teacher perceptions, but as in research on faculty trust, the focus is seldom on the most disenfranchised teachers in the tail of the

Teachers in such an unfortunate position (whether justified or partially justified and partially imagined), may be unlikely to view the scores on evaluative global protocols as malleable, believing them to be irrevocably determined by a hostile party. Agnostic systems are not immune to various sources of bias, but they are less likely to be personally biased towards individual teachers than judgmental systems, or to become more biased over time. If a teacher does make a real change, it won't be discounted. Here, automation will greatly reinforce and compound this effect.[17]

Observational Tools and Their Uses

In the preceding section I argued that an inherent feature of observational tools, the degree of agnosticism, may influence teacher response to feedback provided through that tool through four specific mechanisms. Yet, it is worth emphasizing that even if an inherent feature contributes to or influences teacher response, this is not that same thing as fully *determining* that response. Indeed, some critical determinants of teachers' professional development, including teacher voice, may be most strongly impacted by factors external to the tool itself, including the context and nature of its use. Consider an instructional coaching relationship, where teachers and their coaches are supported by an observational system. Conceptually, the literature on teacher coaching (e.g., Knight, 2011), allows us to see that in such a relationship, a tool might provide the framework or backdrop for coaching, but the constructs that might matter the most in generating teacher learning or other professional growth are actually the normative practices and principles that under-gird that coaching. For example, taking the tool as a given, the amount of teacher voice in the instructional improvement process will be heavily determined by the orientation and actions of the coach, including efforts to identify teachers' own goals and how they want to adapt basic pedagogical principles to their classroom.

Empirically, research by Cherasaro et al. (2016) gives us a high-level view of teacher response to observational evaluation and feedback. Although there are many unknown details about the use of observational tools in this major study, one overall implication is that teacher response is highly variable. This research was in the context of evaluation, concerning the global observation protocols used in district systems of teacher evaluation. Overall, approximately 55% of teachers agreed or strongly agreed the feedback was useful. Despite this very mixed perceptions of overall usefulness, sixty percent or more of teachers reported feedback uptake; planning new instructional activities or approaches, etc. Importantly, while perceptions of overall usefulness and uptake (response to feedback) were related to perceptions of accuracy and evaluator credibility, there was a great deal of unexplained variance in usefulness and uptake. It seems clear that teacher response can't be neatly predicted from basic features of the observational systems in use, both because individuals vary in how they perceive those features, and how the tool is used with teachers likely has a huge effect independent of those features.

Messick (1989) argues that the interpretation and use of scores is fundamental to understanding the validity of those scores as a construct. A logical consequence of that now well-established view (see e.g., Groves et al., 2009; Kane, 2013) is that future research on classroom observation systems should not be restricted to the study of measurement properties, but also engage in the study of use-value, with actual empirical data on protocols in use generated by qualitative and quantitative research (see e.g., Goldring et al., 2015; Kraft & Gilmour, 2016). Use-value will be affected by the nature and context of its use (see above discussion of variation in

distribution, so it is more difficult to estimate what proportion of teachers might hold extreme distrust of observational validity.

[17] Although not narrowly tailored to this topic, research by Logg et al. (2019) suggests that individuals are generally receptive to algorithmic predictions in the modern era.

teacher response in Cheresaro et al., 2016), but I have also argued that the level of agnosticism in an observational system is a "fundamental" feature and may tend to influence teacher response, in general, through specific mechanisms (choice, etc.). Moreover, I do not believe that views of validity that correctly stress interpretation necessarily imply that all understanding of measurement systems is entirely conditional on use. There are two reasons this would not be the most helpful perspective in the study of observational systems. First, inherent properties of observational systems, such as their degree of agnosticism, may have similar effects across multiple (if not all) uses. Second, the mere existence of a given system almost ensures that it will be put to multiple uses. A system's use will rarely be so narrowly tailored that it's validity can be assessed around a single, discrete use.

Conclusion

In this essay I have compared global, judgmental protocols vs. agnostic, fine-grained protocols, and to a lesser extent, considered the role of automation in this comparison. Agnosticism is just one of the important basic features of classroom observation protocols (see Table 1). I have paired agnosticism with grain-size in this essay primarily due to the empirical regularity of this correspondence in the example systems considered, not because the two features are by definition related. However, I have hypothesized that being fine-grained is generative, in a probabilistic, imperfect sense, of agnosticism. My even more basic thesis is that agnosticism itself is an underappreciated concern in observational research. In addition to their obvious necessity in discovery-oriented research, agnostic systems of classroom observation may be useful in various efforts to promote teacher learning. Agnostic systems, I suggest, offer choice, withhold judgement, make room for locally-compensatory practices, and may promote a greater locus of control. Admittedly, it is somewhat counter-intuitive to propose that we might be better able to improve instruction by explicitly avoiding any judgement of its overall effectiveness. But consider an analogy: athletic coaches use stopwatches to gain fine-grained insight into training intervals, but the stopwatch is agnostic about whether on a given day and time, a particular athlete should be doing a more- or less-intense training interval to receive a maximum training benefit. Maybe what educators need are similarly agnostic tools to help leverage their professional judgment and invigorate their professional practice.

Educational researchers should continue to develop and refine global protocols, but also agnostic, fine-grained systems. Researchers working in the automated realm should very purposefully consider measurement agnosticism in goals and methods. A great deal of effort goes into the human-coding side in developing automated methods, and what type of system the human observers use in coding is a "crux" decision that determines and constrains the automation.

References

Assor, A. (2012). Allowing choice and nurturing an inner compass: Educational practices supporting students' need for autonomy. In S. L. Christenson, A. L. Reschly & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 421–440). Springer. https://doi.org/10.1007/978-1-4614-2018-7_20

Aucejo, E. M., Coate, P., Fruehwirth, J., Kelly, S., & Mozenter, Z. (2022). Teacher effectiveness and classroom composition. *The Economic Journal*, *132*, 3047–3064. https://doi.org/10.1093/ej/ueac046

Arehart, J. (1979). Student opportunity to learn related to student achievement of objectives in a probability unit. *Journal of Educational Research*, *72*, 253–269. https://doi.org/10.1080/00220671.1979.10885166

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

Bell, C. A., Drew H. Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012) An argument approach to observation protocol validity, *Educational Assessment*, *17*, 62–87. https://doi.org/10.1080/10627197.2012.715014

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

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*, 1069–1077. https://doi.org/10.1037/0003-066X.41.10.1069

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

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.

Cherasaro, T. L., Brodersen, R. M., Reale, M. L., & Yanoski, D. C. (2016). *Teachers' responses to feedback from evaluators: What feedback characteristics matter?* (REL 2017-190). Regional Educational Laboratory Central.

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

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.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*, 378–387. https://doi.org/10.3102/0013189X16659442

Cooley, W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis*, *2*, 7–25. https://doi.org/10.3102/01623737002001007

Conner, A. M., Singletary, L. M., Smith, R. C., Wagner, P. A., & Francisco, R. T. (2014). Teacher support for collective argumentation: A framework for examining how teachers support students' engagement in mathematical activities. *Educational Studies in Mathematics*, *86*, 401–429. https://doi.org/10.1007/s10649-014-9532-8

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*, 715730. https://doi.org/10.1037/0022-0663.88.4.715

The Danielson Group. (2011). *The framework for teaching evaluation instrument*. Princeton, NJ.

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

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*(4), 531–556. https://doi.org/10.3102/0162373718784205

Dunkin, M. J., & Biddle, B. J. (1974). *The study of teaching*. Holt, Rinehart & Winston.

Dworkin, A. G. (2009). Teacher burnout and teacher resilience: Assessing the impacts of the school accountability movement. In L. J. Saha & A. G. Dworkin (Eds.), *International handbook of* *research on teachers and teaching* (pp. 491–509). Springer. https://doi.org/10.1007/978-0-387-73317-3_32

Foss, R., Diekman, S., Goodwin, A., & Bartley, C. (2003). *Enhancing a norms program to reduce high-risk drinking among first year students*. University of North Carolina-Chapel Hill.

Franklin, R. K., Mitchell, J. O. N., Walters, K. S., Livingston, B., Lineberger, M. B., Putman, C., & Karges-Bone, L. (2018). Using Swivl robotic technology in teacher education preparation: A pilot study. *TechTrends*, *62*, 184–189. https://doi.org/10.1007/s11528-017-0246-5

Furnham, A., & Howard, S. (1993). Measuring locus of control: A critique of general, children's, health- and work-related locus of control questionnaires. *British Journal of Psychology*, *84*, 443–479. https://doi.org/10.1111/j.2044-8295.1993.tb02495.x

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, processes and outcomes* (pp. 109–126). Eliot Werner Publications.

Gamoran, A., & Nystrand, M. (1992). Taking students seriously. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 40–61). Teachers College Press.

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. https://doi.org/10.1177/016146811411600607

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

Good, T. Grouws, D., & Beckerman, T. (1978). Curriculum pacing. Some empirical data in mathematics. *Journal of Curriculum Studies*, *10*, 75–81. https://doi.org/10.1080/0022027780100106

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality.

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.

Grissom, J. A., Blissett, R. S., & Mitani, H. (2018). Evaluating school principals: Supervisor ratings of principal practice and principal job performance. *Educational Evaluation and Policy Analysis*, *40*(3), 446–472. https://doi.org/10.3102/0162373718783883

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. CALDER. https://doi.org/10.3386/w16015

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd edition). Wiley.

Haines, M. (1996). *Social norms approach to preventing binge drinking at colleges and universities*. (Pub. No. ED/OPE/96-18). U.S. Department of Education.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., Robyn, A., Russell, J. L., Naftel, S., & Barney, H. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*. Rand Corporation. https://doi.org/10.7249/MG589

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

Hennessy, S., Howe, C., Mercer, N., & Vrikki, M. (2020). Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture, and Social Interaction*, *25*, 100404. https://doi.org/10.1016/j.lcsi.2020.100404

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

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

Hunter, S. B., & Rodriguez, L. A. (2021). Examining the demands of teacher evaluation: time use, strain and turnover among Tennessee school administrators. *Journal of Educational Administration*. https://doi.org/10.1108/JEA-07-2020-0165

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., & Lai, V. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*. https://doi.org/10.1016/j.tate.2022.103631

Jacoby, A. R., Pattichis, M. S., Celedón-Pattichis, S., & LópezLeiva, C. (2018, April). Context-sensitive human activity classification in collaborative learning environments. In *2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)* (pp. 1–4). IEEE. https://doi.org/10.1109/SSIAI.2018.8470331

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

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

Jensen, E., Dale, M., Donnelly, P., Stone, C., Kelly, S., Godley, A., & D'Mello, S. K. (2020). Toward automated feedback on teacher discourse to enhance teacher learning *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI 2020).* https://doi.org/10.1145/3313831.3376418

Jensen, E., Pugh, S. L., & D'Mello, S. K. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 302–312). https://doi.org/10.1145/3448139.3448168

Jones, N., Bell, C., Qi, Y., Lewis, J., Kirui, D., Stickler, L., & Redash, A. (2021). *Certified to evaluate: Exploring administrator accuracy and beliefs in teacher observation*. ETS Research Report Series. https://doi.org/10.1002/ets2.12316

Juzwik, M., Nystrand, M., Kelly, S., & Sherry, M. (2008). Oral narrative genres as dialogic resources for classroom literature study: A contextualized case study of conversational narrative discussion. *American Educational Research Journal, 45,* 1111–1154. https://doi.org/10.3102/0002831208321444

Kane, M. T. (2013). Validating the interpretations and uses of test scores, *Journal of Educational Measurement*, *50*, 1–7. https://doi.org/10.1111/jedm.12000

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.

Katz, D. (1960). The functional approach to the study of attitudes. *The Public Opinion Quarterly*, *24*, 163–204. https://doi.org/10.1086/266945

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research*, *37*, 434–448. https://doi.org/10.1016/j.ssresearch.2007.08.003

Kelly, S. (2010). The prevalence of developmental instruction in public and Catholic schools. *Teachers College Record*, *112*, 2405–2440. https://doi.org/10.1177/016146811011200906

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). https://doi.org/10.14507/epaa.28.5012

Kelly, S., Guner, G. K., Hunkins, N., & D'Mello, S. K. (in progress). High school English teachers reflect on teacher talk: A user-study of automated classroom observation and feedback.

Kelly, S., Mozenter, Z., Aucejo, E., & Fruehwirth, J. (2020b). School-to-school differences in instructional practice: New descriptive evidence on opportunity to learn. T*eachers College Record*, *122* (11). https://doi.org/10.1177/016146812012201102

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

Knight, J. (2011). What good coaches do: When coaches and teachers interact equally as partners, good things happen. *Educational Leadership*, *69*, 19–22.

Kraft, M. A., & Christian, A. (2019). *In search of high-quality evaluation feedback: An administrator training field experiment*. Ed-Working Paper 19-62. Annenberg Institute at Brown University, Providence, RI.

Kraft, M. A., & Gilmour, A. F. (2016). Can principals promote teacher development as evaluators? A case study of principals' views and experiences. *Educational Administration Quarterly*, *52*, 711–753. https://doi.org/10.1177/0013161X16653445

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

Krosnick, J., & Fabrigar, L. (1997). Designing rating scales for effective measurement in surveys. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., & Trewin, D. (eds.), *Survey measurement and process quality* (pp. 141–164). Wiley. https://doi.org/10.1002/9781118490013.ch6

Lehesvuori, S., Ramnarain, U., & Viiri, J. (2018). Challenging transmission modes of teaching in science classrooms: Enhancing learner-centredness through dialogicity. *Research in Science Education*, *48*, 1049–1069. https://doi.org/10.1007/s11165-016-9598-7

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*, 587–614. https://doi.org/10.3102/01623737211009267

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithmic appreciation: People prefer algorithmic to human judgement. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

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

McCoy, S., Lynam, A., & Kelly, M. (2018). A case for using Swivl for digital observation in an online or blended learning environment. *International Journal of Innovations in Online Education*, *2*. https://doi.org/10.1615/IntJInnovOnlineEdu.2018028647

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*, 5–11. https://doi.org/10.3102/0013189X018002005

MET Project. (2010, October). *The PLATO Protocol for Classroom Observations*. Bill & Melinda Gates Foundation.

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

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). *A composite estimator of effective teaching.* (Tech. Rep.). Bill & Melinda Gates Foundation, Measures of Effective Teaching Project.

Morine-Dershimer,G. (1985). *Talking, listening, and learning in elementary classrooms*. Longman.

Muijs, D., Reynolds, D., Sammons, P., Kyriakides, L., Creemers, B. P., & Teddlie, C. (2018). Assessing individual lessons using a generic teacher observation instrument: How useful is the International System for Teacher Observation and Feedback (ISTOF)? *ZDM*, *50*(3), 395–406. https://doi.org/10.1007/s11858-018-0921-9

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

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

Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in time: Investigating the structure and dynamics of unfolding classroom discourse. *Discourse Processes, 35*(2), 135–198. https://doi.org/10.1207/S15326950DP3502_3

Pace, J. L., & Hemmings, A. (2007). Understanding authority in classrooms: A review of theory, ideology and research. *Review of Educational Research*, *77*, 4–27. https://doi.org/10.3102/003465430298489

Pepper, M. J., Ehlert, M. W., Parsons, E. S., Stahlheber, S. W., & Burns, S. F. (2015). *Educator evaluations in Tennessee: Findings from the 2014 First to the Top survey*. Tennessee Consortium on Research, Evaluation, & Development.

Praetorius, A. K., & Charalambous, C. Y. (2018). Classroom observation frameworks for studying instructional quality: looking back and looking forward. *ZDM, 50*(3), 535–553. https://doi.org/10.1007/s11858-018-0946-0

Price, H. E. (2021). Weathering fluctuations in teacher commitment: Leaders relational failures, with improvement prospects. *Journal of Educational Administration*, *59*, 493–513. https://doi.org/10.1108/JEA-07-2020-0157

Price, H. E. (2012). Principal-teacher interactions: How affective relationships shape principal and teacher attitudes. *Educational Administration Quarterly*, *48*, 39–85. https://doi.org/10.1177/0013161X11417126

Ramakrishnan, A., Zylich, B., Ottmar, E., LoCasale-Crouch, J., & Whitehill, J. (2021). Toward automated classroom observation: Multimodal machine learning to estimate class positive climate and negative climate. *IEEE Transactions on Affective Computing.* https://doi.org/10.1109/TAFFC.2021.3059209

Reznitskaya, A., Anderson, R. C., McNurlen, B., Nguyen-Jahiel, K., Archodidou, A., & Kim, S-O. (2001). Influence of oral discussion on written argument. *Discourse Processes*, *32*, 155–175. https://doi.org/10.1207/S15326950DP3202&3_04

Ridge, B. L., & Lavigne, A. L. (2020). Improving instructional practice through peer observation and feedback. *Education Policy Analysis Archives, 28*(61). https://doi.org/10.14507/epaa.28.5023

Ritchey, F. (1999). *The statistical imagination*. McGraw-Hill.

Rosenshine, B. (1970). Enthusiastic teaching: A research review. *The School Review*, *78*, 499–514. https://doi.org/10.1086/442929

Rosenshine, B. (1971). *Teaching behaviours and student achievement*. National Foundation for Educational Research in England and Wales.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*, 1–28. https://doi.org/10.1037/h0092976

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*, 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Russ, R. S., Sherin, B. L., & Sherin, M. G. (2016). What constitutes teacher learning. In D. H. Gitomer & C. A. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 391–438). AERA. https://doi.org/10.3102/978-0-935302-48-6_6

Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? *Educational Researcher*, *43*, 404–412. https://doi.org/10.3102/0013189X14554450

Schunk, D. H. (1991). Self-efficacy and academic motivation. *Educational Psychologist*, *26*, 207–231. https://doi.org/10.1080/00461520.1991.9653133

Sherif, C. W. (1963). Social categorization as a function of latitude of acceptance and series range. *Journal of Abnormal and Social Psychology*, *67*, 148–56. https://doi.org/10.1037/h0043022

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. Springer. https://doi.org/10.1007/978-1-4614-7089-2

Smith, L. (1979). Task-oriented lessons and student achievement. *Journal of Educational Research*, *73*, 16–19. https://doi.org/10.1080/00220671.1979.10885197

Smith, L., & Land, M. (1981). Low-inference verbal behaviors related to teacher clarity. *Journal of Classroom Interaction*, *17*, 37–42.

Smith, T. W. (1997). Factors relating to misanthropy in contemporary American society. *Social Science Research*, *26*, 170–196. https://doi.org/10.1006/ssre.1997.0592

Tschannen-Moran, M., & Hoy, W. (1998). Trust in schools: A conceptual and empirical analysis. *Journal of Educational Administration*, *36*, 334–352. https://doi.org/10.1108/09578239810211518

Van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50,* 88–95. https://doi.org/10.1016/j.stueduc.2016.08.001

van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J., & Visscher, A. J. (2019). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement*, *30*, 51–67. https://doi.org/10.1080/09243453.2018.1539013

Van Maele, D., & Van Houtte, M. (2009). Faculty trust and organizational school characteristics: An exploration across secondary schools in Flanders. *Educational Administration Quarterly*, *45*, 556–589. https://doi.org/10.1177/0013161X09335141

Watson, G., Youngs, P., van Aswegen, R., & Singh, S. (2021). *Automated classification of elementary instructional objects and activities: Analyzing consistency of manual annotations*. [Paper presentation]. Annual Meeting of the American Educational Research Association (April, online).

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

White, M., Luoto, J. M., Klette, K., & Blikstad-Balas, M. (2021, August 19). Bringing the theory and measurement of teaching into alignment. https://doi.org/10.31219/osf.io/fnhvw

Wilkinson, I. A., Soter, A. O., & Murphy, K. P. (2010). Developing a model of quality talk about literary text. In M. G. McKeown, & L. Kucan (Eds.), *Bringing reading researchers to life: Essays in honor of Isabel L. Beck* (pp. 142–169). Guilford Press.

Zee, M., & Koomen, H. M. Y. (2016). Teacher self-efficacy and its effects on classroom processes, student academic adjustment, and teacher well-being: A synthesis of 40 years of research. *Review of Educational Research*, *86*, 981–1015. https://doi.org/10.3102/0034654315626801

About the Author

Sean Kelly
University of Pittsburgh
spkelly@pitt.edu

Sean Kelly is a Full Professor in the Department of Educational Foundations, Organizations, and Policy at the University of Pittsburgh. He studies the social organization of schools, student engagement, and teacher effectiveness. He currently serves as co-editor of the *American Educational Research Journal.*

education policy analysis archives

Volume 31 Number 7 January 31, 2023 ISSN 1068-2341

About the Editorial Team: https://epaa.asu.edu/ojs/index.php/epaa/about/editorialTeam

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

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