The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/vjxe20

Fine-Grained, Nomination Coding in the Support Domain: Promising Teacher Discourse Measures

Sean Kelly, Hadassah Muthoka, Erin Vines, Stephanie Wormington & Sidney D'Mello

To cite this article: Sean Kelly, Hadassah Muthoka, Erin Vines, Stephanie Wormington & Sidney D'Mello (21 Apr 2024): Fine-Grained, Nomination Coding in the Support Domain: Promising Teacher Discourse Measures, The Journal of Experimental Education, DOI: 10.1080/00220973.2024.2312521

To link to this article: https://doi.org/10.1080/00220973.2024.2312521

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2024.2312521

Routledge
Taylor & Francis Group

Fine-Grained, Nomination Coding in the Support Domain: Promising Teacher Discourse Measures

Sean Kelly[a], Hadassah Muthoka[b], Erin Vines[c], Stephanie Wormington[d], and Sidney D'Mello[e]

[a]School of Education, University of Pittsburgh, Pittsburgh, PA, USA; [b]Motivate Lab, Charlottesville, VA, USA; [c]University of FL, Gainesville, FL, USA; [d]Center for Creative Leadership. One Leadership Place, Greensboro, NC, USA; [e]University of CO Boulder, Boulder, CO, USA

ABSTRACT

In this study, we report results from a novel coding of the Measures of Effective Teaching (MET) Study data that offers evidence on a set of teacher discourse measures in the domain of teacher support including: public praise vs. admonishment, autonomy support vs. controlling language, strategy suggestion vs. lack thereof, and discourse supporting (vs. undermining) learning mindsets. Novel coding of these constructs is paired with extant measures of instruction and achievement in the MET data. Several of the newly coded discourse measures have promising features, including high lesson- and teacher-level variability, and convergent and discriminant validity with existing protocols. We also report possible associations with change in achievement over two years.

KEYWORDS

Classroom observation; discourse; support; engagement; motivation

Introduction

It is widely acknowledged that teacher observations provide a critical external source of information to spur teacher learning and instructional growth (Clarke & Hollingsworth, 2002; Goe et al., 2012; Praetorius & Charalambous, 2018). As a target of observation, prior research has identified teacher support, including socio-emotional supports, broadly construed, as an essential element of effective instruction, particularly in mathematics (Scheerens, 2014; Shernoff, 2013). But what form and focus should teacher observation to inform understanding of teacher support have? To date, much research has focused on global observations of teacher support practices, where observers provide summary ratings of teacher support practices (Hamre et al., 2013; Hill et al., 2008; Li et al., 2020). In this study we focus on a complementary research question: Is supportive discourse a promising candidate for the focus of a *fine-grained* program of research on instructional practice?

To answer this question, we focus on exploring a set of measurement properties in novel data on teacher support practices as enacted *via* discourse in middle-school mathematic classrooms including: overall lesson-to-lesson variability as well as teacher-to-teacher variability in discourse practices, internal consistency, convergent and divergent validity with global protocols, and associations with achievement scores. We envision the measures investigated in this study as broadly useful in the analysis of classroom instruction, including: (1) research on differences across classrooms and schools in typical instructional practices and their effects, as well as (2) professional

CONTACT Sean Kelly spkelly@pitt.edu School of Education, University of Pittsburgh, Pittsburgh, PA, USA.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/00220973.2024.2312521.

development uses in a variety of pre- and in-service contexts where analysis of instruction can support teacher learning.

The support measures investigated in the present study pertain in particular to socio-emotional supports that address students' need for competence (Newmann et al., 1992) as well as social comparison and equity in heterogeneous classrooms (Cohen & Lotan, 1997). We also include learning strategies as an element of support, which may reduce anxiety and failure-avoiding behaviors when confronting difficult material. Socio-emotional and motivational supports may be particularly important in mathematics, a subject in which many students experience achievement anxiety (Wigfield & Meece, 1988) and where levels of interest and enjoyment may be lower than in other subjects (Shernoff et al., 2003). Our focus on mathematics is also motivated by prior research showing a strong relationship between teacher support and student engagement and learning in this domain (Bishop, 2021; Franke et al., 2015; Kelly & Zhang, 2016). Lastly, our focus on observing and coding teacher discourse is motivated by research demonstrating that socio-emotional and motivational supports in mathematics education is revealed and enacted through teacher discourse (e.g., Turner et al., 1998, 2002).

Relative advantages/disadvantages of global vs. fine-grained measures of instruction

Currently, the vast majority of observation-based teacher evaluation and professional development efforts utilize in-person observation with global protocols (Kelly et al., 2020), which provide a summary evaluation of multiple domains of instruction from observers' overall review of classroom processes. In contrast, we argue that a paradigm shift in methods of teacher observation is needed, toward fine-grained measures of instruction (Kelly et al., 2020, Kelly, 2023).

Although some global protocols have special emphases, one strength is that they are generally quite comprehensive of a large set of domains of instructional practice (Praetorius & Charalambous, 2018). For summative assessment, that comprehensiveness is essential, helping to avoid construct under-representation in measurement. Additionally, global protocols are scored over an interval of time, and thus cover all instruction and are not restricted to particular methods/modes of instruction (lecture, question and answer, seatwork) or classroom activity structure (e.g., whole class vs. small group or individualized). Moreover, many protocols are designed to be used flexibly across a variety of subject-matter areas and/or grade levels. In our view, they are well-rooted in the educational sciences of best practices, with content evidence that the domains and indicators constitute effective practice. Overall, global protocols offer a valuable complement to the use of survey reports of instruction and test-based inferences on teaching quality (Kelly et al., 2020). Finally, it is important to note that the overall use of global protocols is not restricted to producing scores used in teacher accountability systems. For example, their use may enhance professionalization by providing teachers and administrators with a shared pedagogical language (Goldring et al., 2015).

Yet, the comprehensiveness, broad applicability, and overall utility in summative assessment of global protocols comes with limitations and tradeoffs (Bell et al., 2014; Gitomer et al., 2014; White, 2018). Most obviously, the large grain-size of measurement means that teachers are provided with feedback for improvement only in general domains, and without much precision. Indeed, global protocol scores tend to cluster in a few modal categories, limiting their ability to guide improvement (Kelly et al., 2020). For example, in the data analyzed by Kelly et al. (2020), on the eight sub-domains of Danielson's Framework for Teaching (FFT), the proportion of teacher observations in the middle two out of four categories (basic, proficient) ranged from a low of 87% (Using Question and Discussion Techniques) to a high of 96% (Communicating with Students). In contrast, fine-grained measures seek to provide feedback very precisely, for example at the level of individual utterances, seconds of time-use, or individual assignments or tasks (e.g.,

Caughlan et al., 2013; Gamoran et al., 1995). As a result, improvements or changes in instruction, even very incremental changes, can be documented.

Another central feature of global protocols is their inherent focus on a continuum of effective practice, that is, the strong valencing of instructional practice as effective or ineffective. This quality precludes more careful inquiry into teachers' curricular and pedagogical emphases, tradeoffs in time use, and instructional practices where competing theoretical perspectives preclude an a-priori judgment of best practice. Stated differently, in focusing solely on a continuum of effective practice, existing global protocols lack the agnostic quality that would allow a program of research to empirically uncover non-obvious, novel relationships between instructional practice and outcomes (Kelly, 2023). Here, we are referring to an agnostic approach to measurement itself rather than hypothesis generation. That is, specifically, that measures are defined without regard to an underlying relationship to effectiveness. For example, in the present study we measure the prevalence of teacher praise and admonishment, and while theories of instruction and learning that stress students' need for competence (Archambault et al., 2010; Newmann et al., 1992) would suggest that the presence of teacher praise is desirable, other research shows more positive motivational outcomes of negative teacher emotions like anger (Taxer & Frenzel, 2020). Thus, our measures simply identify whether praise or admonishment is occurring, not whether it was effective at that moment of instruction (e.g., the admonishment could be directed toward a student bullying another). In contrast, many global observation protocols presume effectiveness as part of the coding. Thus, while much of the praise and admonishment coded here might serve to assign competence to students (see e.g., Cohen & Lotan, 1997) or undermine students' sense of competence, this should not be assumed during coding. Rather, it is recommended that coders simply code occurrences of praise and admonishment, whose effects can later be empirically uncovered.

Considering the use of aggregate data from systems of teacher observations, we argue that this focus on a continuum of effective practice, along with the basic rough, qualitative nature of lesson scoring in global protocols, means they might not be well-suited to assess variation in opportunity to learn specific content, or to document large-scale changes in instructional emphases over-time. Indeed, while many protocols stress multiple domains of instructional practice, factor analytic analyses of the covariance structure of scores suggests fewer instructional constructs are present/referenced in actual use (Aucejo et al., 2022; Humphry & Heldsinger, 2014; Liu et al., 2019; McCaffrey et al., 2015). As an example, consider that standards in mathematics education increasingly stress the incorporation of statistical logic in secondary mathematics classrooms (NGA and CCSSO, 2010). Documenting the changing presence of that content in the curriculum, the allocation of instructional time, and the nature of tasks, assignments, examples, etc., would likely necessitate more fine-grained measures including tools like the Survey of Enacted Curriculum (Porter et al., 2011)' which provides a summary map of curriculum defined by both topical content and cognitive demand.

Yet, by their very nature, fine-grained measures that rely on intensive human coding tend to be difficult and expensive to implement. Thus, in the past, such systems have been primarily used in research settings (e.g., Gamoran & Kelly, 2003; Howe et al., 2019; Murphy et al., 2009; Taylor et al., 2003) and in pre-service teacher preparation (e.g., Juzwik et al., 2013; Kucan, 2009). Today, automated methods of observation and analysis offer the possibility of efficient, fine-grained observation of instruction (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). However, in order to realize the potential of such automated systems, they must be balanced and comprehensive, not narrowly tailored around single constructs, and must be validated on a number of dimensions affecting robust use. This is precisely what we aim to do here with an emphasis on the support domain.

The support domain in mathematics

While educational researchers have not reached full consensus concerning the set of abstract constructs that provide the building blocks for understanding effective instruction, the *support* domain, including emotional and motivational supports, appears in many instructional typologies (Nilsen & Gustafsson, 2016; Scheerens, 2014; Shernoff, 2013). Referencing Shernoff's two-component model of instruction featuring challenge and support, support refers to instructional practices that helps students to meet various challenges inherent to learning novel material and negotiating the social environment of schooling and classrooms.

Supportive classroom instruction, broadly construed, is associated with student engagement (Furrer & Skinner, 2003; Kelly & Zhang, 2016; Wang & Eccles, 2012), academic achievement (Patrick et al., 2007; Perry et al., 2007), and subjective well-being (Suldo et al., 2009; Tennant et al., 2015). Support from teachers may be particularly important in math, where many students experience achievement anxiety (Wigfield & Meece, 1988), as well as for students who might struggle academically (Hamre & Pianta, 2005). It is not easy to characterize the reported *level* of support found in studies of classroom instruction: observational reports find that teachers are generally quite supportive of students (e.g., NICHD-ECCRN, 2002; Perry et al., 2007), while students' own reports are less positive (Klem & Connell, 2004). It is clear however, that teacher support, as experienced by students, is highly *variable* both within and between classrooms (Battistich et al., 1995; Van Houtte & Van Maele, 2012; Schenke et al., 2018), and that variability is very noteworthy for example in the focal math classrooms in the Education Longitudinal Study (Kelly & Zhang, 2016).

What does teacher support look like in Math classrooms? First, consistent with basic models of engagement (Battistich et al., 1995; Marks, 2000; Newmann et al., 1992), support likely entails some universal if abstract principles like valuing student ideas, treating students with respect and fairness, and conveying expectations of learning and success. Yet, mathematics offers some particular engagement challenges, where students' perceived difficulty levels and anxiety are high, and yet interest and enjoyment levels are lower than in other subjects (Goetz et al., 2007; Shernoff, 2013). Thus, research in math instruction has often focused in particular on how teachers aid comprehension and interest (Patrick et al., 2007; Skinner & Belmont, 1993; Sakiz et al., 2012; Turner et al., 1998, 2002). For example, focusing on discourse practices that promote engagement, Turner et al.'s (1998, 2002) observational research emphasizes supportive discourse that helps students negotiate meaning and transfers responsibility to students to learn (e.g., having students "think aloud" as they solve math problems, or explain their reasoning), as well as motivational discourse (e.g., using humor to reduce anxiety about a tough problem, positioning errors as constructive, etc.). Focusing on the role of teacher affective support, Sakiz et al. (2012) found significant associations between perceived teacher affective support and middle school math students' feelings of belonging, self-efficacy, academic enjoyment, and academic effort.

Focal discourse practices in current study

The literature on mathematics instruction offers many instructional constructs related to and comprising teacher support, and even overall instructional frameworks such as the Responsive Classroom Approach and the Mathematics Scan (Ottmar et al., 2015). The Mathematics Scan (M-Scan) was conceptually influential in our work, providing a starting point for developing codes. The M-Scan is a global observation protocol assessing standards-based mathematics instruction on eight domains, including the Mathematical Discourse Community (Berry et al., 2010). However, that tool was designed to produce an ordinal, Likert-scale scoring of over-arching instructional domains/practices, and thus was just a starting point. In contrast, our project goals

entailed utterance-level measurement, prefacing later automated classification in future research (Hunkins et al., 2022). Additionally, given study constraints, it was not possible to identify an exhaustive or fully-comprehensive set of support related constructs. Thus, we inductively identified four paired discourse constructs: **Public Evaluation** of behavior and achievement, or valence of evaluation (Praise vs. Admonishment); **Autonomy Support** (vs. controlling language); **Strategy Suggestion** (vs. lack of strategy); and **Learning Mindset** supportive discourse (vs. mindset undermining discourse).

We treat these constructs as components of a formative, multidimensional model of support (see White et al., 2021 and Kelly et al. (2024) for discussion of formative models), displayed in Figure 1. Note that in Figure 1 the arrows depict the direction of causality being from the measures to the construct, a key feature of formative conceptualizations (Jarvis et al., 2003). In particular, although we view these conceptually as members of a class of support related discourse practices, we do not necessarily view these as reflecting an underlying latent support construct, such that a high internal consistency is implied (Jarvis et al., 2003). Certainly, we do not assume they are all equally well-suited for incorporation into a fine-grained, discourse-based system of instructional measurement, instead we investigate that question here.

Although the process was inductive, we were guided by both general theories of engagement, discourse, and learning in heterogeneous classrooms previously cited, as well as conceptual frameworks in mathematics including the Mathematics Scan in particular (Carpenter & Lehrer, 1999; Langer-Osuna, 2017; Ottmar et al., 2015). The inductive coding approach efficiently identified salient variability in anticipation of eventual automated classification of discourse.

Public evaluation, including praise and admonishment is likely to interact with students' need for competence; where higher ratios of praise to admonishment assign competence and affect student engagement (Cohen & Lotan, 1997; Newmann et al., 1992). In this case, the nature of public evaluation may affect the feedback-efficacy-effort loop, reducing the likelihood of students withholding effort (Schunk & DiBenedetto, 2016). Alternately, the ratio of praise to admonishment may operate more generally, affecting teacher liking and classroom goodwill; but note that one of the few observational studies on related constructs, Schenke et al. (2018), did not find a statistically significant relationship between observed teacher unfairness/unfriendliness and student perceptions of emotional support.

The use of praise, particularly descriptive praise, is a well-established best practice in early-childhood education, and is linked with a variety of positive outcomes for students with emotional and behavioral disorders in particular (Sutherland et al., 2000). The prevalence of praise among older students has also been studied empirically, although it is not clear that sufficiently large-scale, generalizable research has occurred since Brophy's (1981) review to adequately

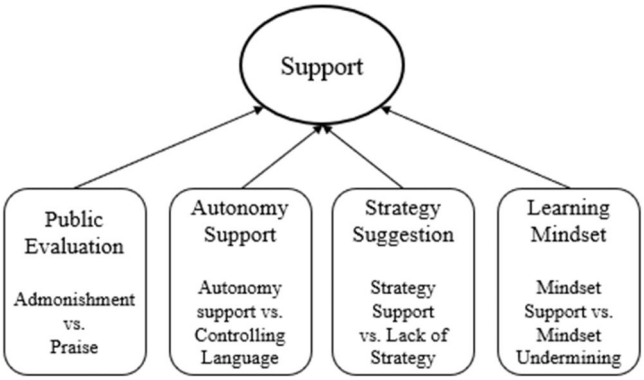

Figure 1. Formative model of support related constructs.

characterize teacher-to-teacher variability in offering praise or its effects (Jenkins et al., 2015). However, intervention studies, based on the finding in some literature that rates of praise are low (e.g., Brophy, 1981), have shown effectiveness in raising the incidence of teacher praise through performance feedback (Hemmeter et al., 2011; Sweigart et al., 2016).

Autonomy support is a central concept in models of teaching and engagement that stress student self-determination (Greene et al., 2004; Stefanou et al., 2004; Su & Reeve, 2011). While we anticipate many elements of autonomy support relate to the nature of classroom tasks and assignments, and the provision of choice in those assignments, teacher discourse, including the absence of controlling language is a key feature of autonomy provision, and part of Reeve and colleagues' successful Autonomy Supportive Intervention Program (ASIP) for teachers (Cheon & Reeve, 2015). In their study of field notes from 106 elementary school classroom observations Bozack et al. (2008) found that opportunities for choice were infrequent and characterized instruction as "teacher centered, teacher directed, and teacher controlled."

Strategy suggestion is a potentially important form of support helping students meet learning challenges that would otherwise cause anxiety or confusion (Shernoff, 2013). For example, strategy suggestion (e.g., concept mapping) features prominently in Guthrie and colleagues' Concept Oriented Reading Instruction (CORI) intervention (Guthrie et al., 2007). A related question is whether the teacher herself is well-guided by a strategic understanding of content learning and common barriers to learning (see e.g., Staples, 2007 case study in Algebra instruction), but we restrict our investigation here to explicit strategy instruction to students.

Of the four constructs, Learning Mindset supportive discourse is the most internally heterogeneous (see coding definition below), as it captures discourse expressing the relevance of mathematics in students' lives (see e.g., Hulleman et al., 2017; Matthews, 2018), but also growth mindsets (Blackwell et al., 2007; Rattan et al., 2012) and social belonging (NRC & IM, 2004; Osterman, 2000; Walton & Cohen, 2007). Although each of these components are conceptually distinct, the discourse coded as learning mindset supportive share a common property/feature in referencing enduring properties of the learning environment and approaches to learning in the classroom, as opposed to the more discrete occurrences of praise/admonishment, etc. Overall, collectively such discourse may affect the assumptions and understandings students hold about themselves, others, and the learning environment that guides behavior (see definition of mindset in Brooks et al., 2012, p. 541). In one empirical example, Bayer et al. (2020) combined these three concepts into a summary RBG index (for Relevance, Belonging, and Growth), finding associations with achievement outcomes in economics classrooms. Additionally, while we appreciate that in other data, researchers might begin by treating these individual components separately, in this study we found they occur only rarely, and even when combined, are the rarest of the four paired constructs (see Table 2).

Importantly, we recognize that these four support constructs are not exhaustive of important discourse moves affecting engagement and learning in mathematics. But how promising are these as discourse constructs (as instantiated here in specific codes, and at the utterance level) as part of a fine-grained program of research on instructional practice? To address this over-arching question we pose the following specific research questions:

RQ1. How prevalent are the measured support-related practices?

RQ2. How much do supportive discourse practices vary from lesson to lesson, and teacher to teacher?

RQ3. Are the measured discourse practices internally consistent?

RQ4. Do the discourse-based measures of support considered have convergent (and divergent) validity with global protocols of instructional quality?

RQ5. Are levels of support associated with student achievement outcomes?

Data and methods

Data

Results are based on a new coding of 156 grade 6–8 math videos, each 15 min in length, from 73 class sections (also 73 teachers) in the longitudinal Measures of Effective Teaching (MET) Study. Thus, we analyze approximately 2.14 video segments per teacher. The MET study was funded by the Bill and Melinda Gates Foundation to investigate ways to identify and develop effective teaching and is available by restricted-use license through the Inter-University Consortium for Political and Social Research (MET Project, 2012, 2014). The MET Study collected data on teachers' instructional practices in six large (but not limited to urban central city schools) school districts over a two-year period from 2009–2010 to 2010–2011. MET offers an exceptionally rich corpus of observational data, random assignment of teachers to students, and other features described in the MET user guide (ICPSR document 34771). At this time, although somewhat dated, these features continue to make MET a particularly valuable corpus of observational data. MET was designed to incorporate data from student assessments, teacher assessments, classroom observations coded with multiple protocols, and student perception data. In addition, a corpus of video data is available for new coding, intended to supplement and extend the original observational coding of instruction.

Newly coded data for this project was merged with existing data on more than 1400 students nested within sample class sections. Our analytic sub-sample of 156 MET videos is a stratified random sample from a set of 446 available high quality audio math section videos in year 2, in which additional ancillary measures are available. Videos were stratified by the dispersion (SD) in overall Tripod scores (a measure of student perceptions of the classroom learning environment), oversampling high- and low-dispersion class sections, in an effort to better capture a wide range of instructional environments. As in the overall MET data, our analytic sample of students and teachers is very diverse, yielding: a student sample approximately 52% male, 47% free- or reduced-price lunch, 33% Latinx, 29% Black; and a teacher sample of approximately 30% male, 39% Black, 10% Latin, with 11.3 years average experience.

Methods

Discourse coding process

To ensure high levels of coder concentration in each video, a random 15-min time interval was selected for coding, making sure disciplinary instruction in mathematics was occurring (as opposed to a classroom interruption, test-taking, "housekeeping" tasks, etc.). Unfortunately, only an approximate matching of new coding efforts to the existing protocol scores is possible in MET for several reasons. First, in some cases, different videos were coded with different extant protocols. Second, only a subset of videos is made available in the restricted use data (about 63% in one calculation with specific sample restrictions). Third, and most problematic, only the entire hour-long video is made available for secondary data analysis, with no indication of start and end times for the segments scored for the extant protocols.

Using a set of initial development videos, the project coding guide was developed by senior study members. We provide this coding guide in an Online Appendix, slightly modified (the original version for coders contained reference to specific MET videos, access procedures, etc.). Thereafter, graduate and undergraduate students were trained to use the protocol and coding reliability was verified at both the utterance and observation level in an initial sample of 15 double-coded observations (see below). We did not have sufficient study scope to conduct a decomposition of variance reliability analysis (see White, 2018), but our key measure of reliability, the observation-level correlation of prevalence rates across raters (i.e., how well did the estimated rate of praise align across the 15 observations, etc.) was so high in most cases, disagreements about

particular utterances did not have strong impact on subsequent analyses. After establishing basic reliability, all remaining files were then coded using an expert-review process, where initial codes were reviewed and corrected by a lead coder (second and third authors), with most corrections entailing additional nominations (see below) missed in the first coding. Although the discourse constructs are somewhat subjective (such that ratings free from classical measurement error might still vary), we do view rater-to-rater differences as reflecting random measurement error and errors in concentration, etc., such that review by lead coders is beneficial. For example, note that the coding of the existing global protocols in MET had very different reliabilities for basic vs. expert coders (see discussion and statistics cited in Kelly et al., 2020). Thus, we believed a lead-coder check was useful in this context.

We used a nomination-coding process to identify teacher utterances exhibiting a given discourse property, as opposed to an exhaustive coding process where all speech is transcribed, segmented, and coded. After nominating an utterance, the teacher speech was entered into an excel database, along with the code and time-stamp (to facilitate future automated analyses). The utterances were naturally and flexibly bounded by the coder to contain the content signaling the nominated support code, which varied in length from as few as 2–4 words (for praise and admonishment) to a median of 24 words for strategy suggestion. This approach yields count data, for example, the number of instances of public praise (per 15-minute interval of time), and ratio data within each of the four discourse domains (e.g., the ratio of praise to admonishment, a relative measure of public evaluation).

Discourse measures

Table 1 provides descriptions and examples of each of the four paired discourse measures.

Public Evaluation of behavior was identified as: public statements of praise, calling out an individual, group of students, or the entire class for ideal or desirable behavior or achievement; or alternatively, public admonishment of a student(s) for disruptive, inappropriate, or undesirable behavior. In these data, empirically, admonishment seldom if ever occurred for incorrect responses or achievement, and instead targeted behavior. Thus, in practice, the two parts of this paired coding pertained to evaluation of different domains of classroom life. Importantly, simple statements of correctness or incorrectness were not coded as evaluation, the evaluation had to be at least minimally elaborated (e.g., for praise, "very good," "nice work," or "great job" are sufficiently elaborated to be coded as praise).

Autonomy Support encompassed a set of discourse practices where students were: provided opportunities for choice; students were expressly directed to engage in independent thinking; alternative solutions were acknowledged and affirmed; opportunities for shared decision-making were offered; or students were allowed to set their own level of challenge. In contrast, controlling language explicitly emphasized students' lack of autonomy and choice; for example, indicating there is only one right answer or method for completing an activity. Of the three forms of autonomy support identified by Stefanou et al. (2004), our coding focused on support for cognitive and procedural autonomy, with less emphasis on organizational autonomy.

Strategy Suggestion in our codebook includes: providing alternative strategies for solving a problem, referencing prior material as a way to understand new material, sharing tips or tricks for remembering material, sharing tips for learning in school overall (e.g., attending tutoring sessions, different study strategies), and providing a tool or scaffolding. Strategy suggestion does not include the simple definition or explanation of concepts and was only coded if it included actual strategies for learning (e.g., employing a process of elimination, searching for context clues, etc.). Lack of strategy was coded only when teacher discourse explicitly emphasized a lack of strategy, as when students were told to "just remember something" or "just figure it out."

Learning Mindset supportive discourse included language related to relevance and personal connections, growth mindsets, and social belonging. Statements that support relevance and

Table 1. Description and examples of focal discourse practices.

	Examples	
Paired discourse construct	Positive	Negative
Public Evaluation		
Description:	*Public praise.* Calling out an individual, group of students, or entire class for ideal or desirable behavior.	*Public admonishment*. Calling out an individual, group of students, or entire class for disruptive, inappropriate, or undesirable behavior.
Examples:	"Beautiful work, ladies, you're rock stars!"	"Excuse me [NAME], for someone who's not feeling well you really are running your mouth"
Autonomy Support		
Description:	*Autonomy*. Providing students with a choice between activities or strategies.	*Controlling language*. Emphasizing the lack of opportunity for autonomy and choice. Often indicates that there is only one right answer or method for completing an activity, and that students do not have agency in how they complete the task.
Examples:	"Now you can do this individually or you can do this in groups." "Okay, and some people have different ways of checking their answer."	"Did it say to do it that way? This is the method we're using. You have to do exactly as it says"
Strategy Suggestion		
Description:	*Strategy suggestion*. Strategy suggestions include sharing techniques, tools, or tips for learning and understanding material.	*Lack of strategy suggestion*. Lack of strategy often refers to "just remembering something," rather than providing a concrete solution for doing so.
Examples:	"How do you usually compare fractions? … You normally turn them into decimals … So if you go to do an experiment in the future and the denominators are not the same, you have two choices. You can go ahead and turn them into the same denominator or you can turn them into decimals."	"Just remember it that way"
Learning Mindset		
Description	*Learning mindset-supportive language.* Language that explicitly supports growth mindset, purpose and relevance, and social belonging.	*Learning Mindset-Undermining Language*. Language that explicitly undermines growth mindset, purpose and relevance, and social belonging.
Examples	"Do you see intersecting lines like everyday? Where?" [personal connections/relevance] "Everyone can learn this, but sometimes you need to try different study strategies to find what works best for you." [growth mindset] "We're a team in this class, so I want you all to work together to figure this out." [social belonging]	"Usually, the way it's going to work is, you fail it the first time, there's a really good chance you're going to fail it the second time."

personal connections might entail connecting course material to real-world activities or topics, connecting to students' interests, families, or communities, or encouraging students to think about their own reasons for valuing material. Statements supporting growth mindsets included highlighting how mistakes lead to learning and growth, conveying the belief that all students can learn and do well in math, and suggesting or praising approaches to learning consistent with a growth mindset. Social belonging related language may underscore students' connection to peers and classmates, that students are cared for by teachers, valued at this school, and acknowledged that it is normal to have feelings of not fitting in. As with other codes, mindset undermining discourse is not just the absence of supportive discourse, but language that actively undermined a growth mindset, relevance, or social belonging.

Table 2. Reliabity and basic properties of nominated teacher discourse measures in the support domain: Counts and Ratio measures. $N = 2{,}818$ utterances in 156 15-min instructional observations. Reliabilities from 15 double-coded observations ($N = 303$).

Inferential Target	Reliability			Prevelence & Variability		Teacher-to-Teacher Variation	Convergent & Divergent Validity[a]		
							Correlations		
Statistics	% Agree	Kappa	Obs-Level Corr	Mean (SD)	Min, Max	ICC (SE)	CLASS_PC	CLASS_SE	MQI
# of Utterances				18.08 (10.38)	9, 104				
Counts									
Praise	96.0	.87	.98	4.52 (3.85)	0, 18	.33 (.10)	.13	.06	.12
Admonishment	92.7	.82	.99	6.56 (7.28)	0, 44	.45 (.09)	−.33***	−.42***	−.06
Autonomy Support	96.0	.58	.87	1.23 (1.20)	0, 5	.25 (.10)	.27***	.26**	−.05
Controlling Language	94.1	.67	.93	1.54 (1.45)	0, 8	.23 (.10)	−.10	−.15	−.08
Strategy Support	94.7	.82	.89	2.31 (1.51)	0, 7	.07 (.11)	.11	.06	.06
Lack of Strategy	99.3	.50	.42	0.17 (.34)	0, 2	.03 (.11)	.01	−.01	.00
Mindset Support	98.7	.88	.97	1.05 (1.09)	0, 6	.43 (.09)	.12	.11	.06
Mindset Undermining	99.3	.87	.96	0.50 (.70)	0, 3	.01 (.11)	−.20*	−.22**	−.16
Cronbach's $\alpha = .10$									
Ratios (positive/(positive + negative))									
Praise Ratio				0.44 (.24)			.24**	.28***	.05
Autonomy Ratio				0.46 (.28)			.29***	.27***	.06
Strategy Ratio				0.94 (.13)			.00	.02	.00
Mindset Ratio				0.82 (.24)			.18*	.23**	.13

*$p < .05$, **$p < .01$, ***$p < .001$.
[a]CLASS_PC (Overall Class Score); CLASS_SE (CLASS Student Engagement Score); MQI (Mathematical Quality of Instruction Score.

Inter-rater reliability

Fifteen double-coded observations yielded 303 nominated codes (see Table 2). Absolute agreement ranged from a low of 92.7% for Admonishment to 99.3% for Lack of Strategy and Mindset Undermining discourse. Taking into account the base rate of expected agreement due to chance (which is very high for low incidence rate codes), Kappa statistics ranged from a low of .50 (Lack of Strategy) to .87–.88 for Praise and Learning Mindset Supportive and Undermining discourse. As in prior work, we place the most emphasis on observation-level correlations between double-coded observations as evidence of reliability (i.e., did the two observers agree on the observation-level prevalence of a given form of discourse), which were above .85 for all measures except lack of strategy suggestions, which has an exceptionally low incidence rate. Thus, with the obvious exception of Lack of Strategy, all measures demonstrated high reliability.

Extant dependent measures

We analyze three scores (see Table 2) from global observation protocols coded as part of the original MET study: The overall CLASS score (CLASS_PC), CLASS student engagement score (CLASS_SE), and holistic overall Mathematical Quality of Instruction score (MQI_OVERALL_HOL). The first two provide potential evidence of convergent validity, with substantial construct overlap with the novel discourse measures we coded. Association with MQI provides potential evidence of divergent validity, as MQI focuses more heavily on teacher accuracy with content and meaning-focused instruction (MET Project, 2012) than support domains. In Table 3, we analyze two types of achievement data, the state administered mathematics achievement tests (available as both pre- and post-test measures), and the Balanced Assessment of Mathematics (BAM), which is available only as a post-test. BAM differs from traditional standardized tests in presenting students with a small number of detailed tasks ($n = 4$–5) which are then scored globally on four dimensions of mathematical thinking (modeling/formulating; transforming/manipulating; inferring/drawing conclusions; and communicating).

Statistics and models

Newly coded measures of instruction were merged with the larger MET database to facilitate a variety of descriptive and inferential analyses. In Table 2, in addition to basic descriptive statistics describing the prevalence and variability of each construct, we report the ICC from a decomposition of variance to measure the extent of teacher-to-teacher variation (as opposed to lesson-to-lesson variation). We also report a measure of overall internal consistency (Cronbach's alpha) of the set of indicators. If each of our constructs were parallel measures of a single construct of "support" and measured with some precision, they would have a high average inter-item correlation (a key component of Cronbach's alpha). Yet, we have conceptualized these measures from a more formative perspective, and as we show, the alpha measure of internal consistency is very low in this case. Next, we report correlations with the CLASS and MQI measures as measures of convergent and divergent validity. Finally, drawing on MET's longitudinal data structure, Table 3 reports a series of three saturated multilevel regression models (unstructured variance-covariance structure of the random effects) of mathematics achievement as a function of the full set of discourse measures (expressed as ratios, e.g., praise/(praise + admonishment)) and the following controls for classroom composition: student age (in years), gender, race/ethnicity (Black, Hispanic, Asian, American Indian, and Other; special education status; gifted status; English language learner status (ELL); and free-lunch status; see Brown et al. (2023) for recent discussion of the robustness of rich-covariate adjustment methods in educational research. Among student background measures, missing data arises from one district that did not report gifted status and another that did not report free-lunch status (Common Core of Data identifiers are not available). The first two models are a regressor-variable and a change-score (see discussion in Kelly & Ye, 2017) specification respectively with state tests, while the third presents association with the BAM assessment.

Results

Table 2 reports basic properties of the four sets of paired teacher support discourse measures. To very briefly summarize the overall findings in Table 2: with the exception of expressions related

Table 3. Association between ratio-form nominated teacher support discourse measures and student math achievement. Multilevel models (STATA xtmixed, mle estimation). $N = 73$ class sections. SE shown in parentheses.

Model	(1)	(2)	(3)
DV	State Math 2011	State Math 2011–2010	BAM
Number of students	1,630	1,635	1,452
Constant	.08 (.34)	−.60 (.34)	2.39
State Math 2010	.61 (.02)***		
State English 2010	.13 (.02)***		
Praise Ratio	.04 (.12)	.09 (.13)	−.23 (.23)
Autonomy Ratio	.17 (.12)	.08 (.11)	.50 (.20)*
Strategy Ratio	.18 (.22)	.36 (.21)	−.35 (.38)
Mindset Ratio	.06 (.12)	.21 (.12)	−.37 (.21)
Age	−.02 (.02)	.01 (.02)	−.15 (.03)***
Male	−.01 (.02)	−.04 (.03)	−0.01 (.04)
Gifted	.15 (.05)**	.00 (.05)	.39 (.08)***
Special Educ	−.06 (.06)	.02 (.06)	−.36 (.09)***
ELL	.00 (.04)	.06 (.05)	−.29 (.07)***
Free Lunch	−.08 (.03)*	−.04 (.03)	−.19 (.05)***
Free Lunch Miss.	−.12 (.08)	−.10 (.08)	.51 (.14)***
Black	−.10 (.04)*	.01 (.04)	−.38 (.06)***
Hispanic	−.05 (.04)	−.05 (.04)	−.11 (.06)
Asian	.17 (.06)**	.06 (.06)	.23 (.09)**
American Indian	−.10 (.15)	.04 (.17)	−.40 (.22)
Race Other	−.07 (.08)	−.05 (.09)	−.10 (.13)

*$p < .05$, **$p < .01$, ***$p < .001$. BAM = Balanced Assessment of Mathematics.

to strategy use, the three remaining paired sets occur with some regularity (RQ1) and vary from lesson to lesson and from teacher to teacher (RQ2). In addition, Public Evaluation, Autonomy Support, and Learning Mindset undermining discourse are correlated in the expected direction with the CLASS observational protocol (demonstrating convergent validity—RQ4), and much less so with MQI, as anticipated evidence of divergent validity (RQ4). Finally, although the teacher discourse constructs investigated here are found regularly in the literature on teacher support and instructional effectiveness, we find very little internal consistency (e.g., average inter-item covariance—RQ3) whatsoever among these measures in this observational data (Cronbach's alpha = .10). Importantly, internal consistency was not implied in our measurement approach—the set of measures is not intended to reflect an underlying central tendency in teacher support. Instead, the measures constitute a set of support-related discourse features that may, formatively, have consequences for the classroom learning environment (see Jarvis et al., 2003 for discussion of internal consistency in formative measurement models).

Returning to the basic prevalence (RQ1) and variability (RQ2), the focal discourse codes in this study occur with regularity—approximately 18 support-related utterances were nominated in each 15 min lesson segment. The sum of positive codes (9.11) was similar to that of negative codes (8.77). Turning to specific codes, both Praise and Admonishment occur frequently and have large total variation (SD), with Admonishment being somewhat more common in these data (Praise Ratio of .44). The prevalence of Public Evaluation is also highly variable from lesson to lesson (SDs of 7.28 and 3.85 for Praise and Admonishment respectively) and teacher to teacher (ICCs of .45 and .33). Autonomy Support, and its antithesis, Controlling Language, occur in about equal measure, vary substantially from lesson to lesson and teacher to teacher, but overall were nominated less frequently than evaluative codes. Learning Mindset supportive discourse occurred much more frequently than Mindset Undermining discourse, which occurs only about once every other observation. Interestingly we see very little variance across teachers in Mindset Undermining discourse; Although relatively rare, almost all teachers occasionally used such discourse. Lack of strategy use occurred even more rarely (not quite once every five lessons), and moreover it was difficult to code reliably (e.g., Kappa of .50). While in theory this form of discourse is recognizable enough (e.g., phrases like, "Just figure it out, please."), in practice it appears to be rare and difficult to identify.

Although prevalence rates are not related, by definition, to the effect of a discourse move, the prevalence rate is related to the likelihood of documenting a substantial, pair-wise difference in any given set of observations (e.g., in comparing two different teachers, or the same teacher's lessons in different periods of the day, or on different days). When prevalence rates are extremely low, any given pair of observations is likely to look exactly the same on that construct, such that a priori, the chance of offering any meaningful comparison whatsoever is reduced. Thus prevalence rates affect the utility of an observation, especially in smaller scale research, professional development efforts, and so on.

Among the three sets of measures where both positive and negative measures are prevalent and reliably coded (i.e., setting aside Strategy Suggestion due to issues with coding Lack of Strategy), all ratio-measures (lower panel of Table 2) show statistically significant associations with scores on the CLASS protocol (ranging from .18 for Mindset Ratio, CLASS_PC to .29 for Autonomy Ratio, CLASS_PC), but null (smaller and non-significant) associations with MQI (discriminant validity). Interestingly, examining the upper panel of results for the count variables, these associations appear to be driven primarily by either the positive or negative expression of the discourse form but not both. That is, Admonishment is associated with CLASS, but not Praise, Autonomy Support but not Controlling Language, and, somewhat unexpectedly given the lack of teacher-to-teacher variance, Learning Mindset Undermining and not Mindset Supportive discourse.

In evaluating the correlations between our novel, fine-grained measures of teacher support and the existing measures of instruction in MET, consider that the low reliability of the existing observational measures in MET is well documented (MET, 2014) and will tend to drive down non-measurement error-corrected correlations such as those in Table 2. Second, the measures being compared in Table 2 are fully independent—common-mode effects present in many studies of the classroom learning environment are known to drive up correlations (Mihaly et al., 2013). Thus, the reported associations with CLASS do provide evidence of convergent validity. In contrast, we did not expect significant associations with MQI because of the different focus that instrument takes—we interpret the comparative lack of correlations with MQI as evidence that common-mode effects are not responsible for the associations we do find with CLASS.

Table 3 explores the relationship between our novel codes for supportive teacher discourse and student achievement in the MET data (RQ5). As general context, consider that it is relatively difficult for researchers to identify teaching practices that consistently increase tested achievement (Kelly et al., 2019). Chetty et al. (2014) report estimates of teacher value-added in mathematics that range from .11 to .16 standard deviations of increase, but that refers to the total effect of all unobservable teacher contributions. The effects of specific teaching practices are often much smaller. For example, Aucejo et al. (2022) find the effects of teaching practice in their analysis of math instruction in the MET data are limited (restricted to higher achieving classrooms). Overall, the present analysis is relatively exploratory and small scale compared to analyses of existing codes in the full MET data.

With that context in mind, associations between supportive teacher discourse and achievement outcomes in Model 1 and 2 are positive, but not statistically significant. Here the dependent measures are standardized, and the independent measures in their raw scale (offering an intuitive 0–1 range, with a SD of about .25 on average, see Table 2). Models considering discourse practices entered one-at-a-time do not differ markedly (recall the low inter-item covariance among this set of measures). Model 3 explores the association with the Balanced Assessment of Mathematics assessment, and here Autonomy Support emerges as a statistically significant correlate of achievement in marked contrast to the other measures of support, which are not only insignificant but also have negative coefficients. Statistically significant results for Autonomy Support hold in BAM models that include pretest controls (the effect declines to .34 (.12 se)), although the effects of many covariates change (free lunch status, etc.). BAM is a very different assessment than the state tests, focusing on extended, complex math tasks where students create a plan, make a decision or solve a problem and then justify their thinking. For example, one middle school task entails responding to an analysis of a confetti-drop in Times Square, including open-ended, constructed responses allowing for creative expression of the amount of confetti dropped. Thus, it is intuitively plausible that supporting student autonomy would be predictive of performance on this task.

Discussion

Much educational research is predicated on the understanding that instructional practices in the support domain, including emotional and motivational supports, are central to effective instruction. Global protocols successfully incorporate this logic, giving educators the ability to provide comprehensive, broadly applicable assessments of teacher support. It seems logical that supportive discourse is a promising candidate for incorporation into fine-grained measures of instruction as well. Beyond that vague promise however, many questions stand between theories of supportive instruction and measures useful in research and professional development. Questions considered here include: prevalence (RQ1), overall lesson-to-lesson variability and teacher-to-teacher variability in particular (RQ2), internal consistency (RQ3), convergent and divergent validity with global

protocols (RQ4), and associations with achievement scores (RQ5). In what follows, we provide a summary of our findings by aligning our four codes with the above research questions.

Interpretive summary of findings

We provide some evidence concerning these questions about teacher discourse practices, using video data to consider all teacher discourse, as opposed to discourse specific to certain activity structures as in some prior research (e.g., Gamoran et al., 1995; Gamoran & Kelly, 2003; Kelly, 2007). We focus on four support constructs, and find that three of these constructs, Evaluation, Autonomy Support, and Learning Mindset Support, appear to hold the most promise for further research. In contrast, fine-grained coding of strategy-related discourse does not seem very useful; such discourse occurs rarely, is difficult to identify reliably in the case of Lack of Strategy, doesn't vary much from teacher-to-teacher, and has little validity evidence. We offer two explanations for this finding. First, strategy use may be a more pervasive aspect of instruction not primarily generated or evinced by the content of individual utterances. For example, assignment prompts and other materials likely perform this function. Second, a little strategy suggestion might go a long way, such that documenting the uniformity/prevalence of it throughout the class session is of less value. Indeed, strategies (and goals) set in previous lessons could carry over to a given lesson, but not be apparent in that lesson's discourse.

Returning to Evaluation, Autonomy Support, and Learning Mindset Support as promising measures, each of these were prevalent (RQ1), varied substantially from lesson to lesson and teacher to teacher (RQ2), were not internally consistent (RQ3), and demonstrated convergent validity with the CLASS protocol (RQ4). Additionally, in these data Autonomy Support was associated with achievement on a more "authentic" math assessment (RQ5). Although again, the achievement models primarily yielded null findings for the other variables. These findings concern central aspects of functionality in discourse measures for instructional research and evaluation, although these features are not necessarily required for certain uses. For example, even measures with relatively low prevalence rates could still be useful in instructional improvement efforts, especially if these measures served as a "tip of the iceberg," being revealing of larger instructional norms, or if when they did occur, their influence was especially strong. Considered as a total set, approximately 17 coded utterances were nominated/identified each lesson. Recall that the prevalence rates reported here refer to 15-min instructional segments, so support related utterances occurred a little more than once per minute.

One of the more surprising findings of this study was the low internal consistency (Cronbach's alpha) of the four paired discourse domains. The nature of Public Evaluation of behavior is almost entirely orthogonal to the level of Autonomy Support, which is in turn orthogonal to Mindset Support, etc. Stated differently, a lesson might, for example, exhibit high levels of Praise but low levels of Autonomy Support, etc. This weak inter-item/measure covariance was not a function of measurement error at the utterance- or lesson-level (except perhaps in the case of Strategy Support, which we found difficult to robustly measure in this framework). Generally, in contrast, global protocols offer high or very high internal consistency, with the various subdomains having high pair-wise correlations and correlations with the scale score (Kelly et al., 2020).

Yet, we argue that the high internal consistency of global protocols is not as desirable as it first appears. This consistency could be caused by training and other underlying teacher quality factors generating a similar competency across various domains. Alternately, it could reflect artificial consistency created by halo-effects in the process of observation/scoring (Liu et al., 2019; McCaffrey et al., 2015). In either case, it means teachers rarely receive information about particular, discrete areas of teaching in need of improvement. We posit that the low internal consistency found in these fine-grained measures is evidence that this measurement approach avoids halo effects,

where raters are not carefully discriminating among scores on sub-domains. It may also be evidence that, especially for research purposes, teacher observation protocols should be weary of overly-abstracted coding of teacher practices (i.e., where abstraction exacerbates the halo effects). Conceptually, it may even be evidence for a formative as opposed to reflective view of teaching practice (see discussion in Kelly & Zhang, 2016; Kelly et al., 2020; White et al., 2021). Overall, we do not view the low internal consistency found here as a fundamental limitation in fine-grained research. To the contrary, it shows that each construct is measured independently. It does however mean that an individual construct (e.g., the prevalence of praise) should not be taken, unto itself, as an accurate representation of the overall socio-emotional environment of the classroom.

Limitations & future work

We conclude with a discussion of limitations and further development prospects of fine-grained discourse-based measures of instruction, and by distinguishing between implications for use in research and practice. Concerning limitations and further development, we emphasize several design features which might be altered in future study. First, critical features of this study were the nomination coding approach, and the focus on discourse to begin with. As a result, the study does not offer and test an exhaustive model of all forms of support. The nomination approach streamlined coding, allowing us to efficiently code more discourse constructs, for more videos, than a variable-based coding of every utterance. Nevertheless, it may be less precise than more intensive coding, or make the data less useful in validating automated coding methods.

The inductive approach we took to identifying codes steered us toward a focus on at least somewhat prevalent practices. We could not document potentially conceptually important practices that currently occur only rarely (because such practices would not have caught our attention as salient). In the case of public evaluation, this resulted in praise and admonishment codes capturing different domains, where praise was almost always achievement related and admonishment was almost always behaviorally related. In the case of learning mindset-supportive discourse, we combined occurrences of conceptually distinct language related to growth mindsets, purpose and relevance, and social belonging. In subsequent research these domains could be separated. Regarding autonomy support, we did not focus on or document organizational autonomy support (e.g., student involvement in setting classroom rules, due dates for assignments, etc.)

Second, while we are relatively confident that the three prevalent/reliable paired measures vary substantially from teacher-to-teacher, a study featuring more intensive study (more observations) for each participant, would sharpen this understanding. In all cases the standard errors in the ICCs (a function of the sample size/nesting) are quite large, such that the relative degree of teacher-to-teacher variation is estimated only imprecisely. Thus, understanding of these constructs would benefit from a larger corpus of data. In particular, given the prevalence rate of Learning Mindset related discourse, and the low teacher-to-teacher variability of Mindset Undermining discourse, a more substantial amount of observational data is needed to robustly study these rare-event constructs.

Finally, because MET did not visually roster/map classrooms with student ids, we could not match teacher utterances to individual participants (as in e.g., Kelly, 2008), which limits inferences about variability in students' experiences within classrooms. Considering the many research design possibilities in fine-grained research on instruction, together with the inherent complexity of conceptual models of learning, the continued collection of high-quality instructional data like MET, that can then be flexibly analyzed, will be critical to advancing the educational sciences.

Use-value of measures

Do the implications of this study for use of such measures in research differ from use in practice? As previously stated, one of the major advantages of global protocols is their comprehensive nature (even as some protocols emphasize socio-emotional supports or other major domains of instruction). Especially for use in teacher evaluation, being quite comprehensive is important, as it avoids construct under-representation in assessing "instructional quality" writ large. Evaluation has not been our goal in this study, so comprehensiveness is of lesser concern. For use in professional development, the fact that scores on various sub-domains of global protocols are not as separable in practice as the structure of the protocol would make it appear (Humphry & Heldsinger, 2014; Kelly et al., 2020; Liu et al., 2019; McCaffrey et al., 2015) suggests that teachers are not actually getting the domain-specific feedback they think they are getting. In contrast, this study presented measures that were largely orthogonal, which may be desirable in a feedback system.

Yet, outcomes of even narrowly-tailored use of a system of instructional observation are not necessarily tightly coupled with measurement properties. For example, even if aspects of measurement are flawed, observational tools might drive school improvement in practice settings. Goldring et al. (2015) find that even as global protocol scores are often unreliable and clustered in the middle of the scale, principals report that the observational frameworks themselves both focus teacher attention and reflection on important domains of instruction and enhance professionalization through a shared technical language. Similarly, we must assess the various uses and use-value of fine-grained observational systems, rather than relying on predictions from measurement models. Such systems have been successfully used to document differences in opportunity to learn across classrooms and schools, although given the labor costs of human coded efforts, only rarely so (see e.g., Gamoran et al., 1995; Gamoran & Kelly, 2003).[1] Fine-grained systems have also been put to productive use in teacher education and professional development settings (Caughlan et al., 2013; Lehesvuori et al., 2017; Reznitskaya & Wilkinson, 2021; Sherry et al., 2018). Yet, less is known about less structured use of such tools, for example, how teachers might use fine-grained, automated systems. As fine-grained measures are evaluated across a range of uses, their possibilities and limitations will be brought into sharper focus.

Note

1. The use of fine-grained instructional measures was more common in the process-product era of research, see summaries in Brophy (1986) and elsewhere.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by a grant from the Student Experience Research Network (formerly Mindset Scholars Network), K-12 Teachers and Classrooms Research Portfolio.

References

Archambault, I., Eccles, J. S., & Vida, M. N. (2010). Ability self-concepts and subjective value in literacy: Joint trajectories from grades 1 through 12. *Journal of Educational Psychology*, *102*(4), 804–816. https://doi.org/10.1037/a0021075

Aucejo, E., Coate, P., Fruehwirth, J., Kelly, S., & Mozenter, Z. (2022). Teacher effectiveness and classroom composition: Understanding match effects in the classroom. *Economic Journal*, *132*(648), 3047–3064. https://doi.org/10.1093/ej/ueac046

Battistich, V., Solomon, D., Kim, D., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance. *American Educational Research Journal*, *32*(3), 627–658. https://doi.org/10.3102/00028312032003627

Bayer, A., Bhanot, S. P., Bronchetti, E. T., & O'Connell, S. A. (2020). Diagnosing the learning environment for diverse students in introductory economics. *AEA Papers and Proceedings*, *110*, 294–298. https://doi.org/10.1257/pandp.20201051

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 Kane, T., Kerr, K., Pianta, R. (Eds.), *Designing Teacher Evaluation Systems: New Guidance from the Measures of Effective Teaching Project* (pp. 50–97). Jossey-Bass.

Berry, R. Q., III, Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. (2010). *The Mathematics Scan (M-Scan): A Measure of Mathematics Instructional Quality*. Unpublished measure, University of Virginia.

Bishop, J. P. (2021). Responsiveness and intellectual work: Features of mathematics classroom discourse related to student achievement. *Journal of the Learning Sciences*, *30*(3), 466–508. https://doi.org/10.1080/10508406.2021.1922413

Blackwell, L. A., Trzesniewski, K. H., & Dweck, C. S. (2007). Theories of intelligence and achievement across the junior high school transition: A longitudinal study and intervention. *Child Development*, *78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Bozack, A. R., Vega, R., McCaslin, M., & Good, T. L. (2008). Teacher support of student autonomy in comprehensive school reform classrooms. *Teachers College Record: The Voice of Scholarship in Education*, *110*(11), 2389–2407. https://doi.org/10.1177/016146810811001110

Brooks, R., Brooks, S., & Goldstein, S. (2012). The power of mindsets: Nurturing engagement, motivation, and resilience in students. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 541–562). Springer.

Brophy, J. (1981). Teacher praise: A functional analysis. *Review of Educational Research*, *51*(1), 5–32. https://doi.org/10.3102/00346543051001005

Brophy, J. (1986). Teacher influences on student achievement. *American Psychologist*, *41*(10), 1069–1077. https://doi.org/10.1037/0003-066X.41.10.1069

Brown, S., Song, M., Cook, T. D., & Garet, M. S. (2023). Combining a local comparison group, a pretest measure, and rich covariates: How well do they collectively reduce bias in nonequivalent comparison group designs? *American Educational Research Journal*, *60*(1), 141–182. https://doi.org/10.3102/00028312221136565

Carpenter, T. P., & Lehrer, R. (1999). Teaching and learning mathematics with understanding. In E. Fennema & T. Romberg (Eds.), Mathematics classrooms that promote understanding (pp. 19–32). Mahwah, NJ: Lawrence Erlbaum.

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.

Cheon, S. H., & Reeve, J. (2015). A classroom-based intervention to help teachers decrease students' amotivation. *Contemporary Educational Psychology*, *40*, 99–111. https://doi.org/10.1016/j.cedpsych.2014.06.004

Chetty, R., Friedman, J.N. & Rocko, J. E. (2014). Measuring the impacts of teachers: Evaluating bias in teacher value-added estimates. *The American Economic Review, 104*, 2593–2632.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education*, *18*(8), 947–967. https://doi.org/10.1016/S0742-051X(02)00053-7

Cohen, E. G., & Lotan, R. A. (1997). *Working for equity in heterogeneous classrooms: Sociological theory in practice*. Teachers College Press.

Franke, M. L., Turrou, A. C., Webb, N. M., Ing, M., Wong, J., Shin, N., & Fernandez, C. (2015). Student engagement with others' mathematical ideas: The role of teacher invitation and support moves. *Elementary School Journal*, *116*(1), 126–148. https://doi.org/10.1086/683174

Franklin, R. K., Mitchell, J. O., Walters, K. S., Livingston, B., Lineberger, M. B., Putman, C., Yarborough, R., & Karges-Bone, L. (2018). Using Swivl robotic technology in teacher education preparation: A pilot study. *TechTrends, 62*(2), 184–189. https://doi.org/10.1007/s11528-017-0246-5

Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, *95*(1), 148–162. https://doi.org/10.1037/0022-0663.95.1.148

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

Gamoran, A., Nystrand, M., Berends, M., & Lepore, P. C. (1995). An organizational analysis of the effects of ability grouping. *American Educational Research Journal*, *32*(4), 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: The Voice of Scholarship in Education*, *116*(6), 1–32. https://doi.org/10.1177/016146811411600607

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.

Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between-and within-domain relations of students' academic emotions. *Journal of Educational Psychology*, *99*(4), 715–733. https://doi.org/10.1037/0022-0663.99.4.715

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*(2), 96–104. https://doi.org/10.3102/0013189X15575031

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivations. *Contemporary Educational Psychology*, *29*(4), 462–482. https://doi.org/10.1016/j.cedpsych.2004.01.006

Guthrie, J. T., Hoa, L. W., Wigfield, A., Tonks, S. M., Humenick, N. M., & Littles, E. (2007). Reading motivation and reading comprehension growth in the later elementary years. *Contemporary Educational Psychology*, *32*(3), 282–313. https://doi.org/10.1016/j.cedpsych.2006.05.004

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, *76*(5), 949–967. https://doi.org/10.1111/j.1467-8624.2005.00889.x

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. *Elementary School Journal*, *113*(4), 461–487. https://doi.org/10.1086/669616

Hemmeter, M. L., Snyder, P., Kinder, K., & Artman, K. (2011). Impact of performance feedback delivered via electronic mail on preschool teachers' use of descriptive praise. *Early Childhood Research Quarterly*, *26*(1), 96–109. https://doi.org/10.1016/j.ecresq.2010.05.004

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*(4), 430–511. https://doi.org/10.1080/07370000802177235

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4–5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Hulleman, C. S., Kosovich, J. J., Barron, K. E., & Daniel, D. B. (2017). Making connections: Replicating and extending the utility value intervention in the classroom. *Journal of Educational Psychology*, *109*(3), 387–404. https://doi.org/10.1037/edu0000146

Humphry, S. M., & Heldsinger, S. A. (2014). Common structural design features of rubrics may represent a threat to validity. *Educational Researcher*, *43*(5), 253–263. https://doi.org/10.3102/0013189X14542154

Hunkins, N., Kelly, S., D'Mello, S. (2022). Beautiful work, you're rock stars!": Teacher analytics to uncover discourse that supports or undermines student motivation, identity, and belonging in classrooms. *12th International Conference on Learning Analytics and Knowledge (LAK'22, Online, March 21-25, 2022)*.

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussion in mathematics classrooms: Using personalized automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, 103631. 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.

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*(2), 199–218. https://doi.org/10.1086/376806

Jenkins, L. N., Floress, M. T., & Reinke, W. (2015). Rates and types of teacher praise: A review and future directions. *Psychology in the Schools*, *52*(5), 463–476. https://doi.org/10.1002/pits.21835

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

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

Kelly, S. (2007). Classroom discourse and the distribution of student engagement. *Social Psychology of Education*, *10*(3), 331–352. https://doi.org/10.1007/s11218-007-9024-0

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research*, *37*(2), 434–448. https://doi.org/10.1016/j.ssresearch.2007.08.003

Kelly, S. (2023). Agnosticism in instructional observation systems. *Education Policy Analysis Archives*, *31*(7), 1–26. https://doi.org/10.14507/epaa.31.7493

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

Kelly, S., Guner, G., Hunkins, N., & D'Mello, S. (2024). High school English teachers reflect on their talk: A user-study of automated feedback with the Teacher Talk Tool.

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*(7), 451–464. https://doi.org/10.3102/0013189X18785613

Kelly, S., Pogodzinski, B., & Zhang, Y. (2019). Teaching quality. In B. Schneider & G. Saw, (Eds.). *Handbook of the sociology of education in the 21st century* (pp. 275–296). Springer.

Kelly, S., & Ye, F. (2017). Accounting for the relationship between initial status and growth in regression models. *Journal of Experimental Education*, *85*(3), 353–375. https://doi.org/10.1080/00220973.2016.1160357

Kelly, S., & Zhang, Y. (2016). Teacher support and engagement in math and science: Evidence from HSLS. *High School Journal*, *99*(2), 141–165. https://doi.org/10.1353/hsj.2016.0005

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, *74*(7), 262–273. https://doi.org/10.1111/j.1746-1561.2004.tb08283.x

Kucan, L. (2009). Engaging teachers in investigating their teaching as a linguistic enterprise: The case of comprehension instruction in the context of discussion. *Reading Psychology*, *30*(1), 51–87. https://doi.org/10.1080/02702710802274770

Langer-Osuna, J. M. (2017). Authority, identity, and collaborative mathematics. *Journal for Research in Mathematics Education*, *48*(3), 237–247. https://doi.org/10.5951/jresematheduc.48.3.0237

Lehesvuori, S., Hähkiöniemi, M., Jokiranta, K., Nieminen, P., Hiltunen, J., & Viiri, J. (2017). Enhancing dialogic argumentation in mathematics and science. *Studia Paedagogica*, *22*(4), 55–76. https://doi.org/10.5817/SP2017-4-4

Li, H., Liu, J., & Hunter, C. V. (2020). A meta-analysis of the factor structure of the Classroom Assessment Scoring System (CLASS). *Journal of Experimental Education*, *88*(2), 265–287. https://doi.org/10.1080/00220973.2018.1551184

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*(1), 61–95. https://doi.org/10.1007/s11092-018-09291-3

Liu, J., & Cohen, J. (2021). Measuring teaching practices at scale: A novel application of text-as-data methods. *Educational Evaluation and Policy Analysis*, *43*(4), 587–614. https://doi.org/10.3102/01623737211009267

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in elementary, middle, and high school years. *American Educational Research Journal*, *37*(1), 153–184. https://doi.org/10.3102/00028312037001153

Matthews, J. S. (2018). When am I ever going to use this in the real world? Cognitive flexibility and urban adolescents' negotiation of the value of mathematics. *Journal of Educational Psychology*, *110*(5), 726–746. https://doi.org/10.1037/edu0000242

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

MET Project. (2012). *Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation.

MET Project. (2014). *Measures of effective teaching: 1 – Study information, user guide*. Bill and Melinda Gates Foundation.

MET Project. (2014). *Measures of effective teaching: 1 – Study information, observation measures report*. Bill and Melinda Gates Foundation.

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.

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

National Institute for Child Health and Human Development and Early Child Care Research Network. (2002). The relation of global first-grade classroom environment to structural classroom features and teacher and student behaviors. *Elementary School Journal*, *102*, 367–387.

National Research Council and Institute of Medicine. (2004). *Engaging schools: Fostering high school students' motivation to learn*. National Academies Press.

Newmann, F. M., Wehlage, G. G., & Lamborn, S. D. (1992). The significance and sources of student engagement. In F. M. Newmann (Ed.), *Student engagement and achievement in American secondary schools* (pp. 11–39). Teachers College Press.

NGA and CCSSO. (2010). Appendix A: Research supporting key elements of the standards. In Common core state standards for English language arts and literacy in history/social studies, science, and technical subjects. National Governors Association and Council of Chief State School Officers.

Nilsen, T., & Gustafsson, J. E. (2016). *Teacher quality, instructional quality, and student outcomes: Relationships across countries, cohorts, and time* (Vol. 2). Springer.

Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research*, *70*(3), 323–367. https://doi.org/10.3102/00346543070003323

Ottmar, E. R., Rimm-Kaufman, S. E., Larsen, R. A., & Berry, R. Q. (2015). Mathematical knowledge for teaching, standards-based mathematics teaching practices, and student achievement in the context of the responsive classroom approach. *American Educational Research Journal*, *52*(4), 787–821. https://doi.org/10.3102/0002831215579484

Patrick, H., Ryan, A. M., & Kaplan, A. (2007). Early adolescents' perceptions of the classroom social environment, motivational beliefs, and engagement. *Journal of Educational Psychology*, *99*(1), 83–98. https://doi.org/10.1037/0022-0663.99.1.83

Perry, K. E., Donohue, K. M., & Weinstein, R. S. (2007). Teaching practices and the promotion of achievement and adjustment in first grade. *Journal of School Psychology*, *45*(3), 269–292. https://doi.org/10.1016/j.jsp.2007.02.005

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new U.S. intended curriculum. *Educational Researcher*, *40*(3), 103–116. https://doi.org/10.3102/0013189X11405038

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

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*, *14*(1), 664–669.

Rattan, A., Good, C., & Dweck, C. (2012). Its ok—not everyone can be good at math: Instructors with an entity theory comfort (and demotivate) students. *Journal of Experimental Social Psychology*, *48*(3), 731–737. https://doi.org/10.1016/j.jesp.2011.12.012

Reznitskaya, A., & Wilkinson, I. A. (2021). The argumentation rating tool: Assessing and supporting teacher facilitation and student argumentation during text-based discussions. *Teaching and Teacher Education*, *106*, 103464. https://doi.org/10.1016/j.tate.2021.103464

Sakiz, G., Pape, S. J., & Hoy, A. W. (2012). Does perceived teacher affective support matter for middle school students in mathematics classrooms? *Journal of School Psychology*, *50*(2), 235–255. https://doi.org/10.1016/j.jsp.2011.10.005

Scheerens, J. (2014). School, teaching, and system effectiveness: Some comments on three state-of-the-art reviews. *School Effectiveness and School Improvement*, *25*(2), 282–290. https://doi.org/10.1080/09243453.2014.885453

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction*, *55*, 13–21. https://doi.org/10.1016/j.learninstruc.2018.02.003

Schunk, D. H., & DiBenedetto, M. K. (2016). Self-efficacy theory in education. In *Handbook of motivation at school* (pp. 34–54). Routledge.

Shernoff, D. J. (2013). *Optimal learning environments to promote student engagement*. Springer.

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly*, *18*, 158–176.

Sherry, M. B., Messier-Jones, L. M., & Morales, J. (2018). Positioning in prospective secondary English teachers' annotations of teaching videos. *English Teaching: Practice & Critique*, *17*(3), 152–167. https://doi.org/10.1108/ETPC-11-2017-0154

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, *85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Staples, M. (2007). Supporting whole-class collaborative inquiry in a secondary mathematics classroom. *Cognition and Instruction*, *25*(2–3), 161–217. https://doi.org/10.1080/07370000701301125

Stefanou, C. R., Perencevich, K. C., DiCintio, M., & Turner, J. C. (2004). Supporting autonomy in the classroom: Ways teachers encourage student decision making and ownership. *Educational Psychologist*, *39*(2), 97–110. https://doi.org/10.1207/s15326985ep3902_2

Su, Y. L., & Reeve, J. (2011). A meta-analysis of the effectiveness of intervention programs designed to support autonomy. *Educational Psychology Review*, *23*(1), 159–188. https://doi.org/10.1007/s10648-010-9142-7

Suldo, S. M., Friedrich, A. A., White, T., Farmer, J., Minch, D., & Michalowski, J. (2009). Teacher support and adolescents' subjective well-being: A mixed-methods investigation. *School Psychology Review*, *38*(1), 67–85. https://doi.org/10.1080/02796015.2009.12087850

Sutherland, K. S., Wehby, J. H., & Copeland, S. R. (2000). Effect of varying rates of behavior-specific praise on the on-task behavior of students with EBD. *Journal of Emotional and Behavioral Disorders*, *8*(1), 2–8. https://doi.org/10.1177/106342660000800101

Sweigart, C. A., Collins, L. W., Evanovich, L. L., & Cook, S. C. (2016). An evaluation of the evidence base for performance feedback to improve teacher praise using CEC's quality indicators. *Education and Treatment of Children*, *39*(4), 419–444. https://doi.org/10.1353/etc.2016.0019

Taxer, J. L., & Frenzel, A. C. (2020). Brief research report: The message behind teacher emotions. *Journal of Experimental Education*, *88*(4), 595–604. https://doi.org/10.1080/00220973.2019.1588699

Taylor, B. M., Pearson, D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms: The influence of teacher practices that encourage cognitive engagement in literacy learning. *Elementary School Journal*, *104*(1), 3–28. https://doi.org/10.1086/499740

Tennant, J. E., Demaray, M. K., Malecki, C. K., Terry, M. N., Clary, M., & Elzinga, N. (2015). Students' ratings of teacher support and academic and social–emotional well-being. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association*, *30*(4), 494–512. https://doi.org/10.1037/spq0000106

Turner, J. C., Meyer, D. K., Anderman, E. M., Midgley, C., Gheen, M., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, *94*(1), 88–106. https://doi.org/10.1037/0022-0663.94.1.88

Turner, J. C., Meyer, D. K., Cox, K. C., Logan, C., Dicintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology*, *90*(4), 730–745. https://doi.org/10.1037/0022-0663.90.4.730

Van Houtte, M., & Van Maele, D. (2012). Students' sense of belonging in technical/vocational schools versus academic schools: The mediating role of faculty trust in students. *Teachers College Record: The Voice of Scholarship in Education*, *114*(7), 1–36. https://doi.org/10.1177/016146811211400706

Walton, G. M., & Cohen, G. L. (2007). A question of belonging: Race, social fit, and achievement. *Journal of Personality and Social Psychology*, *92*(1), 82–96. https://doi.org/10.1037/0022-3514.92.1.82

Wang, M., & Eccles, J. S. (2012). Social support matters: Longitudinal effects of social support on three dimensions of school engagement from middle to high school. *Child Development*, *83*(3), 877–895. https://doi.org/10.1111/j.1467-8624.2012.01745.x

Watson, G., Youngs, P., van Aswegen, R., Singh, S. (2021). Automated classification of elementary instructional objects and activities: Analyzing consistency of manual annotations. *The 2021 Annual Meeting of the American Educational Research Association (April, Online Pandemic Accommodation)*.

White, M. C. (2018). Rater performance standards for classroom observation measures. *Educational Researcher*, *47*(8), 492–501. https://doi.org/10.3102/0013189X18785623

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

Wigfield, A., & Meece, J. L. (1988). Math anxiety in elementary and secondary school students. *Journal of Educational Psychology*, *80*(2), 210–216. https://doi.org/10.1037/0022-0663.80.2.210