EdInstruments

Considerations for Use of Teaching Quality Measures

Matthew Ronfeldt & Matthew Truwit
University of Michigan

As part of the larger library of instruments, EdInstruments catalogs teaching assessments for use in research and practice. These instruments are organized within Schooling > Teaching on the EdInstruments website. This memo establishes a framework of considerations for evaluating and using measures of teaching quality before applying these considerations to the broad sample of 51 such instruments on EdInstruments. The memo is organized into three sections: (1) an effort to conceptualize teaching quality and establish a framework of considerations for its measurement; (2) a brief introduction to the EdInstruments site and its measures of teaching quality; and (3) an assessment of the field of measuring teaching quality, by applying the framework established to the measures available. While we do not endorse individual instruments, we provide examples of instruments for various uses. Users can navigate the EdInstruments website to explore additional options.

Part I: What is teaching quality? How do we measure it?

What is teaching quality?

Pinning down an exact definition of teaching quality is an elusive and perhaps impossible task; as David Berliner concisely states, "Defining *quality* always requires value judgments about which disagreements abound" (2005, p. 206). Teaching is a complex and multidimensional process that can be carried out in many different ways; when coupled with the "contested nature of quality, [one might wonder] is there any sure way to tease out the characteristics and properties of quality teaching?" (Fenstermacher & Richardson, 2005, p. 186).

Despite the philosophical challenges inherent in defining teaching quality, many educational researchers have dedicated their careers to the task. These efforts have only served to reinforce the breadth of what teaching quality can encompass. For example, based on reviews of classroom observation systems, Bell et al. (2019) identify a vast number of dimensions of teaching quality, including a safe and stimulating classroom climate, productive classroom management, high levels of student involvement and motivation, clear explanation of subject matter, rich, precise, and accurate representations of subject matter, deep cognitive activation, cyclical assessment for learning, highly differentiated instruction, and the explicit modeling and scaffolding of learning and self-regulation strategies for students. Meanwhile, Goe et al. (2008) comb research literature, policy documents, reports, and more to construct a five-point definition of effective teachers, who "have high expectations for all students …, contribute to positive academic, attitudinal, and social outcomes for students …, use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence …, contribute to the development of classrooms and schools that value diversity and civic-mindedness, … [and] collaborate with other[s] … to ensure student success, particularly the success of students with special needs and those at high risk for failure" (p. 8). Darling-Hammond (2021) studies the standards for practice in five different high-achieving countries to consider how definitions of teaching quality vary around the world, finding common conceptualization as a wide-ranging knowledge base of not only content and pedagogy (or the learning process) but also the diverse social, emotional, and academic needs of students in order to effectively respond to their individual trajectories for learning; moreover, she notes that "the framework for defining teaching quality has expanded and defines an increasingly evidence- and inquiry-based conception of practice …, as well as one that is increasingly child-centered and focused on concerns for equity and multiculturalism" (p. 296).

In short, although there is near universal agreement that teaching quality matters, there is far from universal agreement as to what exactly it entails. The clearest conclusion is that teaching quality is multifaceted, contextually dependent, and connected to student learning. However, even this last point is not without a difference of opinion. While many have adopted a parsimonious operationalization of teaching quality that is narrowly based on the learning of students, this empirical turn has been met with a wide range of legitimate critiques (Goe et al., 2008). Fenstermacher and Richardson (2005) cleverly resolve this tension by providing a helpful distillation of *quality teaching* into two related concepts—good teaching and successful teaching. *Good teaching* entails meeting thresholds of adequacy in the task sense, performing the logical, psychological, and moral acts of teaching in a way that "comports with morally defensible and rationally sound principles of instructional practice. *Successful teaching* is [that which] yields the intended learning" (p. 189). In their conceptualization, one can enact *good teaching* without yielding intended learning (i.e., *successful teaching*); likewise, *successful teaching* need not be *good* (e.g., effectively teaching a child how to pickpocket). *Quality teaching* occurs only when both conditions are met.

We do not bring these examples to bear with the goal of settling on a conclusive definition of teaching quality. Instead, we illuminate the field's diversity of opinions in order to

underscore the importance of considering one's own perspective as to what teaching quality means before embarking on its measurement. A clear conceptualization of teaching quality—or an explicit bounding of teaching quality to a well-defined dimension—is the first step in deciding how to measure it. Ultimately, we agree with Gitomer and Bell (2013), who state, "Although reasonable people disagree about what distinguishes high-quality teaching, it is important to identify clearly the constructs that comprise teaching quality and how those constructs may be understood relative to the measures used" (p. 8).

How do we measure teaching quality?[1]

Many—but not all—of the challenges in measuring teaching quality stem directly from the difficulty of defining it. According to Goe et al. (2008), "measuring teacher effectiveness has remained elusive in part because of ongoing debate about what an effective teacher is and does" (p. 2). However, this ambiguity notwithstanding, a number of efforts to develop evaluative frameworks have provided a way to generate, select, and/or judge instruments designed to measure teaching quality (Bell et al., 2019; Gitomer, 2019; Goe et al., 2008). We find the framework introduced by Gitomer (2019) particularly helpful for considering different approaches to evaluating teaching quality. Drawing on modern validity theory, Gitomer judges measures based on three broad facets: (1) the *domain* of teaching quality, (2) the *nature of evidence collected and analyzed*, and (3) the process of *interpreting evidence to support inferences*. We draw heavily on, further unpack, and slightly modify (informed by the others cited above) this framework below.

What is the domain of teaching quality to be evaluated?

Any consideration of a measurement of teaching should begin by asking "what aspects of teaching are of interest, how is the domain conceptualized, and what is its theoretical basis?" (Gitomer, 2019, p. 69). As we elaborated more deeply above, clearly defining the construct or specifying one or more well-conceptualized dimensions is the natural first step of measuring teaching quality. This decision is not one to be made lightly; how comprehensive of a conception of teaching one embraces involves a tradeoff between measuring teaching more exhaustively but superficially versus more deeply but narrowly. Although all of the dimensions of teaching quality play a meaningful role in students' learning and development, any single

[1] One of the foremost challenges of measuring *teaching* quality stems from the need to distinguish it from *teacher* quality. Much of the political debate uses the language of "teacher quality" or "teacher effectiveness"; however, this terminology is indicative of a conversation that overlooks the "situational factors that may have a strong bearing on the quality of the teaching practices we see" (Kennedy, 2010, p. 591). While the goal of measurement may be to make inferences or claims about the quality of a given teacher, the practice of measurement is inevitably "intertwined with critical contextual features, such as the curriculum, school leadership, district policies, and so on" (Gitomer & Bell, p. 9). The quality of teaching that is measured is surely a function of the quality of a given teacher but is also a function of the available resources and opportunities, the surrounding context, and the learners involved (Fensterbacher & Richardson, 2005). As a result, given the near impossibility of disentangling the individual teacher from all additional contextual effects, we mostly focus our discussion on measuring *teaching* quality without the intention of attributing this entirely to the teacher, though we do discuss the affordances and constraints of various types of measures in this regard.

approach attempting to capture the entire complex domain of teaching quality will almost inevitably fail to tap all aspects of a multifaceted construct (Gitomer, 2019). Choosing how much of the concept of teaching quality to try to measure is a reflection of both one's individual conceptualization of teaching quality and the intended use of the measure.

Additionally, recent work has begun to consider the ways in which aspects of teaching quality are uniquely influenced by subject matter or discipline (Shulman, 1998; Hill, Schilling, & Ball, 2004). While most of this work has focused on the various forms of knowledge necessary for high-quality teaching (Ball et al., 2008), there is also evidence of discipline-specific teacher practices and their particular efficacy for promoting student learning (Seidel & Shavelson, 2007). As a result, a related choice in how to measure teaching quality involves deciding whether to examine a general or discipline-specific conception of teaching quality.

One additional consideration involving which domains of teaching to measure has largely gone unexplored in prior frameworks—that is, whether and how to conceptualize teaching quality as a practice concerned with culture, race, equity, and diversity. As both learning theory and the field of education have increasingly embraced the need to explicitly attend to racial/ethnic, gender, religious, socioeconomic, and other identities—as well as the corresponding issues of power and privilege—the conceptualization and measurement of teaching quality have placed a priority on dimensions like multiculturalism, social justice, culturally relevant/sustaining pedagogies, and anti-racist instruction. Though some argue that high-quality teaching practices exist agnostic of students' identities—as suggested rhetorically by the title of Ladson-Billing's seminal piece "But That's Just Good Teaching!" (1995a)—many have compellingly articulated the need for teachers to "link principles of learning with deep understanding of (and appreciation for) culture" (Ladson-Billings, 2014, p. 77). We argue that the consideration of which aspects of teaching to measure and how to conceptualize them should also include explicit attention to the degree to which such teaching is informed by culture and oriented to equity.

What is the nature of evidence collected and analyzed to provide insight about teaching quality?

After determining the domain of teaching quality to be considered, it is necessary to consider how it can be operationalized through the collection of evidence. We begin with the nature of that evidence before turning to the practical and logistical aspects necessary to consider during its collection and analysis.

<u>Nature of the Evidence</u>

In their chapter from the *APA Handbook of Testing and Assessment in Psychology*, Gitomer and Bell (2013) propose a conceptualization of teaching quality that is "interactional and constructive", meaning "[w]ithin specific teaching and learning contexts, teachers and students construct a set of interactions that is defined as teaching quality" (p. 9). More importantly for our purposes, in order to measure teaching quality, the authors encourage the

examination of six broad kinds of evidence: the *knowledge, practices,* and *beliefs* of *teachers* and of *students*. We find this framework especially useful for thinking about the many kinds of evidence that can be collected as a measure of teaching quality.

The first three kinds of evidence all originate with the teacher. As such, they are perhaps more indicative of *good teaching*, as defined by Fenstermacher and Richardson (2005). *Teacher practices* include the particular pedagogies, methods, and moves enacted by teachers, typically captured through classroom observation protocols. However, these practices must also be guided by *teacher knowledge*, including a deep and contextualized understanding of students, content, and pedagogy (Shulman, 1998), and *teacher beliefs*, such as the necessary mindset that all students can learn and/or have cultural knowledge that should be leveraged as assets in teaching (Ladson-Billings, 1995b; Lee, 2007; Moll et al., 1992)—as well as the sense of self-efficacy that one has the capacity to do so. These latter two are often measured on questionnaires completed by teachers themselves.

However, as Fenstermacher and Richardson (2005) lay out, *good teaching* on the part of the teacher is not always the same as *quality teaching*, which also involves learning on the part of the students (or *successful teaching*). As such, measurement of teaching quality—whether focused broadly on the entire domain of teaching or specifically on a given dimension—should incorporate a focus not only on the teacher but also on their students. The *student practices* highlighted by Gitomer and Bell (2013) as potential measures of teaching quality include specific classroom-level behaviors, often collected in direct interaction with the *teacher practices* measured via observation systems as described above, but also encompass broader trends like course-taking patterns and graduation rates which can be captured using administrative data. Similarly, *student knowledge*—and particularly knowledge of content presumably learned from teachers—can include both classroom measures and more standardized assessments, with a heavy focus over the past two decades on value-added to student achievement measures (VAMs) which leverage the difference between predicted and actual student scores on state or district assessments to infer a teacher's level of instructional effectiveness. Finally, questionnaires that collect data on *student beliefs* may reflect teaching quality through either teachers' cultivation of students' mindsets, efficacy, and critical thinking or students' own perceptions of the quality of teaching and learning experienced in their classrooms.

Each of these six kinds of evidence has the potential to capture important dimensions of teaching quality; however, each is also unavoidably limited in scope and characterized by unique measurement challenges that will likely result in underrepresentation. For example, a measure that focuses on *teacher practices* without considering *teacher knowledge* might misattribute the rationale for a particular decision, conflating the circumstances of the situation with an indication of the quality of teaching (Kennedy, 2010); meanwhile, focusing on teacher knowledge alone is insufficient, given the evidence that strong content knowledge does not indicate skill at facilitating the learning of that content with students (Grossman, 1989). Similarly, skillful pedagogy and rich professional knowledge are also unlikely to be enough if teachers do not possess the appropriate *beliefs* and mindsets, though a sense of teaching

efficacy and a belief that all students can learn will obviously not suffice without the pedagogical skill to embody them in practice. Any comprehensive assessment of teaching quality must then attend to all three—teacher practices, knowledge, and beliefs; otherwise, it is critical to carefully consider the ways in which measuring only one or two might fall short of representing the complex, multifaceted nature of teaching quality in full.

Similarly, the conceptualization and measurement of teaching quality cannot be limited to a focus on teachers and their teaching exclusively but also must incorporate students and their learning. Finding a set of teachers to excel on measures of their practices, knowledge, and beliefs would not be enough to conclude that they are high-quality teachers if, over time, no learning occurs among their students (Fenstermacher & Richardson, 2005; Gitomer & Bell, 2013). Yet, at the same time, equating teaching quality strictly with impacts on student learning fails to consider the ways that contextual factors in classrooms, schools, and communities may impede or facilitate the manifestation of good teaching as learning. In fact, for learning to occur, Fenstermacher and Richardson (2005) argue that good teaching is but one of four ingredients—the others being the opportunity to teach/learn, a supportive social surround for teaching and learning, and a willingness and effort on the part of the learner. These other three ingredients suggest that broader contextual factors impact whether and how *good teaching* (and learning) manifests into *quality teaching*.

Practical Considerations

Beyond the nature of the evidence it intends to capture, there are also practical considerations necessary for designing, selecting, or evaluating an instrument. These considerations are manifold and heavily dependent on the context in which the measure will be used. For example, instruments will vary in grain size (Bell et al., 2019); that is, measures will differ in the number of items, scale (e.g., a binary scale of present/absent vs. a Likert scale of intensity or frequency), the time required for completion, and the granularity of the teaching instance being measured (e.g., from a segment of a lesson to a broader representation of teaching over time). Additionally, measurement inescapably has cost demands. While some instruments are open access, others require a fee for their use. Costs—both monetary and in terms of time—can also stem from the need for adaptation (of an instrument to a new context, language, or domain) and/or training (of observers or raters to ensure the reliability of their scores). Finally, a number of logistical considerations—paper or digital, in English or in another language, and so on—may play an important role in measuring teaching quality.

How is the evidence interpreted and evaluated to support inferences about teaching quality?

A final[2] consideration in measuring teaching quality involves assessing each instrument's validity (i.e., does an instrument actually measure (only) what it purports to measure?) and

[2] Many would argue that considerations of validity and purpose of use should be the foremost, and not the final, consideration. In many cases, we agree. However, given that the purpose of this document is to provide guidance in creating, choosing, and evaluating measures of teaching quality, we discuss concerns of domain and evidence first.

reliability (i.e., does an instrument produce consistent scores?). Modern validity theory has established that no instrument is valid in and of itself; instead, it may only be demonstrated as valid for a specific use (Messick, 1989). We consider use not only to be the specific purpose for collecting the measure (e.g., research, professional development, high-stakes summative assessment, etc.) but also where along the developmental continuum a measure falls—specifically, how long ago it was established, how widely it has been adopted and in how broad of contexts (Bell et al., 2019). With a purpose or use in mind, it is then possible to collect empirical evidence to build a strong argument for the validity of an instrument when used in a particular context (Kane, 2006). This empirical evidence can come in a wide variety of forms, including (but not limited to):

1. *Content validity*, wherein the items or elements of a measure are vetted by a team of field experts to ensure they are relevant and representative of the construct of interest
2. *Predictive* or *criterion validity*, wherein the scores from an instrument are illustrated to positively (*convergent*) or negatively (*divergent*) associate with those from other validated measures in anticipated or theoretically consonant ways
3. *Construct validity*, wherein the items or elements of an instrument are demonstrated to measure the intended construct with accuracy
4. *Measurement invariance*, wherein an instrument is shown to produce consistently valid scores across subgroups and settings
5. *Reliability*[3], wherein the scores from an instrument are reproduced consistently across settings, instances, and scorers

When evaluating or selecting an instrument, it is crucial to build an argument drawing on these and other forms of validity. An instrument that—again, when used for a specific purpose or in a particular way—is supported by a wide body of empirical evidence illustrating that it consistently and accurately measures a given conceptualization of teaching quality (or one of its many domains) would make a far more attractive option than one backed by limited psychometric research.

[3] Though typically considered a separate construct from validity, we include reliability as a form of empirical validity evidence because a measure cannot be valid if it is unreliable; that is, it cannot measure (only) what it is supposed to measure if it produces inconsistent scores.

Part II: Measures of teaching on the EdInstruments website

Within its larger collection of instruments, the EdInstruments website has aggregated 51 different measures of teaching that range widely in both their operationalizations of teaching quality and their details of implementation. The measures were selected by the EdInstruments team using expert advice and scholarly database searches. Each measure has a page complete with a brief description, details about the content of the instrument, information regarding its administration and access, and links to scholarly work regarding both its use and its psychometric properties. All instruments can be found within the category of "Schooling" and the subcategory of "Teaching."

We caution that the measures on the EdInstruments site are not a comprehensive or exhaustive accounting of the entire pool in use today, nor are they necessarily likely to be representative of that pool. However, we consider the collection a useful cross-section of common instruments that can still provide insight into current trends in how teaching quality is commonly conceptualized and measured today. While our findings can only truly generalize to the collection on the EdInstruments site, we hope that our conclusions will have two outcomes: (1) that they will spur other scholars of measurement to expand this collection by contributing instruments that will make the collection more representative of the tools currently in use and (2) that other scholars will consider whether our conclusions apply to other collections of instruments beyond those included in this collection to see whether and how they translate to the state of measuring teaching quality more generally.

Part III: What can we learn from the EdInstruments site about the state of measuring teaching?

In this section, we apply the framework outlined above to the 51 measures of teaching included on the EdInstruments site, making claims about the larger state of measuring teaching when appropriate. Within each subsection, we first document any noteworthy trends (or "main effects"); we then consider how these trends might overlap or intersect with patterns identified in other subsections (as "interactions").

Domain of Teaching Quality

As teaching quality is a multifaceted and complex construct, attempts at its operationalization naturally take many different approaches. Given that any single measure will inherently fall short of representing all of the facets of the entire construct of teaching quality, some choose to focus narrowly on developing a rich representation of only a particular component while others aim for a broad and more comprehensive operationalization at the expense of depth, time, cost, or some other dimension. Each of the measures on the EdInstruments website uniquely conceptualizes teaching quality—or a specific dimension of

teaching quality—in a way that balances breadth and depth. This choice directly reflects how scores from the instrument are intended to be used.

A few instruments very specifically target only a particular facet of teaching quality—in some cases, students' sense of classroom belonging or teachers' conceptions of empathy; in others, a subcomponent of content knowledge for teaching focused entirely on likely student science misconceptions. These provide a deep, detailed operationalization of one of the many aspects of teaching while eschewing the rest. Other instruments broaden their conceptualizations of teaching quality to cover a wider range of teacher practices or beliefs but circumscribe these to those that fit a particular domain. For example, some measures consider all student-teacher interactions in the classroom, but only concentrate on the social and relational dimensions of these processes; others examine all instructional moves and tasks but with an eye only toward evaluating their rigor and cognitive demand. Most instruments choose an optimum somewhere closer to the middle, seeking to measure the instructional practices of teachers along as many dimensions of teaching quality as possible within reasonable logistical constraints; these instruments often portray themselves as global measures of classroom instruction or pedagogy. However, even these "global" measures, often by drawing evidence heavily from teachers' practices and/or questionnaire responses, miss out on important elements of teaching quality, such as students' perceptions of classroom culture or out-of-classroom teacher practices, e.g., those related to parent and community engagement. Consequently, there are another few measures on the site that combine various types of evidence and operationalizations to support even more exhaustive conceptualizations of teaching quality, but these are the instruments that typically require the most time and training to use, and the extent to which they richly cover all of the facets of teaching quality is still up for debate.

What matters the most here is that the EdInstruments site offers a diverse selection of instruments that range across this spectrum of breadth and depth. While the modal instrument is a purportedly "global" measure that falls toward the middle of this continuum, there are some that treat teaching quality more broadly and others more narrowly. We emphasize again that evaluating, adopting, and implementing an instrument requires a careful consideration of one's own understanding of teaching quality and—most importantly—the intended use of the scores provided by the instrument.

The domain of teaching quality measured by each instrument may differ in ways beyond scope and specificity. With respect to discipline, just less than half of the measures on the EdInstruments website (n = 24) are intended for teachers in all subject areas. Among those that are subject-specific, most target mathematics (n = 15) or science (n = 10) teachers, though two each are designed for ELA and special education teachers. The website lists no subject-specific instruments for social studies/history, foreign language, or physical education teachers.

More than one-quarter (n = 15) of the instruments on the EdInstruments site are fully or partially focused on measuring dimensions of teaching related to multiculturalism, racial equity, or culturally relevant teaching/pedagogy. Given numerous and relatively recent critiques of the

lack of focus on these dimensions of teaching, we view their strong representation on the EdInstruments site as promising.

Layering these two lenses together, only three of the measures categorized with a domain of "Culturally Responsive Teaching" on the EdInstruments site are subject-specific (*n* = 2 for math; *n* = 1 for special education). While a few additional examples of math- and science-focused measures of teaching for equity, social justice, and diversity exist, this remains an area where well-developed instruments are relatively uncommon (see Chang and Cochran-Smith, 2022 for an assessment of the field of such measures used in preservice teacher preparation).

Nature of the Evidence

When considering the six kinds of evidence of teaching quality outlined by Gitomer and Bell (2008), we make two overarching observations of the measures included on the EdInstruments site: (1) these measures disproportionately draw evidence from teachers rather than from students, and (2) among the former, they focus overwhelmingly on teacher practice and beliefs rather than knowledge.

Regarding (1), of the 51 measures, only four have students (rather than teachers or outside observers) as the main respondents.[4] As such, we believe that the instruments on the EdInstruments site may privilege the measurement of *good teaching* over *successful* or *quality teaching*. Traditionally, *successful teaching* has most often been demonstrated through the use of student test scores and value-added to student achievement measures (VAMs). However, research has shown that student surveys demonstrate significant variation among classrooms—within and between schools—and are predictive of gains in student achievement (Ferguson, 2000). Furthermore, studies have illustrated how portfolios of artifacts and related materials incorporating student work can also effectively characterize classroom instruction (Borko et al., 2007).

In terms of intersecting trends, it is worth acknowledging that half (*n* = 2) of the instruments explicitly drawing on evidence of student practices focus on measuring culturally responsive teaching. However, while this is perhaps suggestive of a higher tendency among equity-oriented measures—compared to other measures of classroom instruction—to incorporate student perspectives, these instruments have still been critiqued for failing to center evidence from students enough (Chang & Cochran-Smith, 2022).

Regarding (2), 21 of the measures are observation rubrics that focus on teacher practices, typically completed by outside observers (e.g., principals, administrators, coaches), while 23 are questionnaires completed by teachers and meant to capture their perception and

[4] Some, though not all, of the measures completed by outside observers that are predominantly focused on classroom instruction naturally include a partial focus on student practices (e.g., CLASS). However, the amount of evidence drawn from student practices in these measures usually pales in magnitude to that drawn from teacher practices.

recall regarding various kinds of beliefs, attitudes, and efficacy—including beliefs about teaching and students, attitudinal beliefs, self-efficacy, self-perceptions about their use of certain practices and skill at using them (e.g., culturally relevant pedagogy), efficacy beliefs, mindsets, and feelings. By contrast, only thirteen of the measures focus on teacher knowledge (three of which are different grade band versions of the same measure—DTAMS), typically targeting content knowledge and various forms of pedagogical content knowledge.

Interestingly, ten of the knowledge measures are intended for use with math and science teachers while three focus on multicultural or culturally responsive teaching; none focus on non-STEM content areas or subject-general teaching. Observational measures of teachers' classroom instructional practices and questionnaires involving teacher beliefs were equity-oriented, subject-specific, and subject-general in roughly equal parts.

Practical Considerations

The measures on the EdInstruments site appear practical for a wide variety of contexts and settings. In terms of grade level, instruments for middle school teachers are most common (*n* = 38), followed by K-8 (*n* = 33), high school (*n* = 31), pre-Kindergarten (*n* = 10), post-secondary (*n* = 6), and, finally, early-childhood teachers (<3 years; *n* = 3). While nearly all instruments were used with in-service teachers, a significant portion also were employed in studies measuring the teaching quality of pre-service teachers. All instruments on the EdInstruments website are offered in English, though some have also been translated to Spanish as well as a variety of other languages (e.g., Turkish, Greek, Korean, Chinese, Brazilian, and Portuguese).

Regarding considerations of access and cost, the majority of measures on the EdInstruments website do not require training for use (*n* = 32); however, many (*n* = 19)—and nearly all of the observational measures of classroom instruction—do. The majority of instruments posted are open access (*n* = 32), though many still are not (*n* = 19). A final set of practical considerations involve the granularity of the instrument.[5] The measures on the EdInstruments site range widely in the number of items they comprise, from as few as six to as many as 237, though the questionnaires—completed mostly by teachers themselves—tend to be shorter than the observation rubrics used by external evaluators. Most (*n* = 45) measures placed items on a graded Likert scale, though the number of response options on these scales ranged across instruments from three to seven or more. Furthermore, a few measures instead (or in addition) scored items—typically those focused on teacher and/or student practices/behaviors—in more unique ways, including as a binary checklist (e.g., present-absent), a percentage of time present, a tallied observed rate per minute, or even a rich, open-ended description. Lastly, for observational measures in particular, granularity can also vary in terms of the length or amount of time spent in the classroom observing teaching practice. Some observational instruments focus narrowly on a 15-minute assessment window, while

[5] While this is related to the comprehensiveness or exhaustiveness of a measure's conceptualization of teaching as described above, it is also a technical and logistical factor in evaluating, adopting, and implementing an instrument.

others require multiple half- or even full-day visits before assigning any ratings of teaching quality. We do not observe any obvious differences in the logistical granularity (i.e., length, time, etc.) of instruments across conceptualizations of teaching as a general practice, a subject-specific practice, or an equity-informed/culturally responsive practice.

We find it valuable to briefly discuss how the subset of instruments used specifically in early childhood education differ from those used in K-12 settings. These measures are fewer in number, more recent in their development, and less heterogeneous in their conceptualizations of teaching and the forms of evidence they collect. Compared to the rest of the measures on the EdInstruments site, these early childhood measures were disproportionately observational instruments of teacher practices, with very few questionnaires drawing on teachers' knowledge or beliefs. Moreover, these measures largely conceptualized teacher practices as generalist instruction, rarely[6] centering subject-specific or equity-oriented forms of early-childhood teaching. These patterns suggest a need in the field of measurement for more diverse conceptualizations of early childhood teaching quality, including more conceptualizations that are subject-specific and equity-oriented.

Validity Evidence

In examining the research literature of the measures included on the EdInstruments website, we find substantial variability in the amount, comprehensiveness, and quality of psychometric evidence for each instrument, with some backed by hundreds of research studies published in high-quality journals and others cited only in a single, non-peer-reviewed study or report. We identified six instruments with at least two studies providing evidence for each of the five psychometric categories we considered—content validity, construct validity, convergent validity, measurement invariance, and reliability. Though we do not deeply investigate the rigor or findings of these studies, we take this large amount of diverse psychometric inquiry as evidence of the extent to which these instruments have been deeply explored along dimensions of validity. On the other hand, many instruments have minimal psychometric support from research. One-third of instruments (n = 17) have evidence from one study or fewer across all five psychometric categories, including one instrument with no research evidence in any of these categories.

These differences tend to reflect the prevalence and purpose of each instrument as well as its current stage along a continuum of development and use. For example, the measures supported by more robust psychometric evidence tend to be those developed many years ago and whose use has expanded beyond strictly for research to also include professional development or coaching, program evaluation, and even summative or high-stakes assessment. By contrast, these latter 17 instruments often were developed as part of a single research study and were rarely used in subsequent research or practice. The prevalence of these "one-off" instruments raises some concern that the educational research community may be (1)

[6] *COEMET* is one exception that has a subject-specific focus on mathematics teaching. The Early Childhood Ecology Scale is an equity-oriented exception that emphasizes culturally relevant pedagogy.

prioritizing the creation of new instruments over producing rigorous psychometric evidence for existing and/or adapted instruments and (2) failing to adequately share, collaboratively adapt, and psychometrically evaluate existing measures. That said, we are hopeful that the large number of instruments that are now open access, as well as this EdInstruments site which promises to facilitate the sharing of instruments and corresponding evidence, will help address these concerns.

When examining the specific kinds of psychometric evidence that were most and least common across instruments, we noticed that research tended to privilege forms of construct, predictive, and (to a lesser extent) content validity over evidence of measurement invariance or reliability.[7] 42 instruments had at least one or more sources of evidence for construct validity, typically in the form of fit statistics and factor loadings obtained via factor analysis. 40 instruments also had at least one or more sources of evidence for predictive validity, typically in the form of correlations between the focal measures and other measures thought to proxy for related constructs. 38 instruments were supported by at least one or more studies providing evidence of content validity, though, compared to construct and predictive validity, it was far more common for instruments to have only a single source of evidence (versus multiple sources). Content validity evidence typically involved instrument review by an expert panel (e.g., content experts or teachers) to provide feedback on how comprehensively the intended constructs were represented, interviews with respondents to talk through perceptions and understandings of items, or findings of improved teacher performance after participation in professional development focused on the content of the instrument (compared to other teachers). For almost half of the instruments on the EdInstruments website, we were unable to find any research providing reliability across time points, raters, or forms (n = 23) or evidence of measurement invariance across respondent subgroups (n = 22); when available, the former most often came in the form of inter-rater reliability (typically for observational rubrics) and a few instances of test-retest or item reliability, while the latter took many forms including examining measure performance across gender/racial identities, cultural and national groups, inservice versus preservice teachers, and certification route or area.

We briefly explore the extent to which the research and validity evidence varies across certain subsets of measures on the EdInstruments site. For example, we find that the amount and rigor of validity evidence in support of math-specific measures of teaching is particularly high; this is unsurprising given the large amount of work on content knowledge for teaching and pedagogical content knowledge that originated with many math education researchers (e.g., Ball et al., 2008). Perhaps more surprising is the degree of psychometric evidence backing the more recent—and less diverse—instruments of early childhood teaching quality; though efforts at measuring early childhood education instruction may currently trail the rest of the field in terms of quantity and diversity, existing measures appear to offer relatively substantial validity and reliability evidence. On the other hand, instruments that focused specifically on

[7] Many studies that conducted factor analyses as evidence of construct validity also presented evidence of reliability in terms of internal consistency (e.g., Cronbach's alpha). However, we restrict our consideration of reliability here to consistency across time points, raters, or measures rather than internal consistency across a set of items.

equity-oriented instruction informed by race and culture tended to have fewer studies offering psychometric support, a finding that is reinforced by the call for more content validity checks involving family and community members in Chang and Cochran-Smith (2022). Finally, the few measures drawing mostly on student evidence were also relatively under-supported by psychometric research, underscoring the need in the field for greater attention to how the perspectives and practices of learners can best be incorporated into efforts to measure teaching quality.

References

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389-407.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement, 30(*1), 3–29.

Berliner, D. C. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education*, *56*(3), 205–213.

Borko, H., Stecher, B., & Kuffner, K. (2007). *Using artifacts to characterize reform-oriented instruction: The SCOOP Notebook and rating guide* (CSE Technical Report No. 707). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA.

Chang, W. C. & Cochran-Smith, M. (2022). Learning to teach for equity, social justice, and/or diversity: Do the measures measure up? *Journal of Teacher Education*. Advance online publication. https://doi.org/10.1177/00224871221075284.

Darling-Hammond, L. (2021). Defining teaching quality around the world. *European Journal of Teacher Education*, 44(3), 295-308.

Fenstermacher, G. & Richardson, V. (2005). On making determinations of quality teaching. *Teachers College Record, 107*(1), 186-213.

Ferguson, R. F. (2000). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24-28.

Gitomer, D. H. (2019). Evaluating instructional quality. *School Effectiveness and School Improvement*, *30*(1), 68–78.

Gitomer, D. H., & Bell, C. A. (2013). Evaluating teaching and teachers. In *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* (pp. 415-444). American Psychological Association.

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

Grossman, P. (1989). Learning to teach without teacher education. *Teachers College Record*, 91(2), 191-208.

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11-30.

Kane, M. 2006. Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). New York, NY: American Council on Education and Macmillan.

Kennedy, M. M. (2010). Attribution error and the quest for teacher quality. *Educational Researcher*, *39*(8), 591–598.

Ladson-Billings, G. (1995a). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice*, *34*(3), 159–165.

Ladson-Billings, G. (1995b). Toward a theory of culturally relevant pedagogy. *American Educational Research Journal*, *32*(3), 465-491.

Ladson-Billings, G. (2014). Culturally relevant pedagogy 2.0: A.k.a. The remix. *Harvard Educational Review*, *84*(1), 74–84.

Lee, C. (2007). Culture, literacy, and learning: Taking bloom in the midst of the whirlwind. New York: Teachers College Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Moll, L. C., Amanti, C., Neff, D., & Gonzalez, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory into Practice*, *31*(2), 132-141.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*(4), 454–499.

Shulman, L. (1998). Theory, practice and the education of professionals. *The Elementary School Journal*, 98(5), 511-526.

Suggested citation:

Ronfeldt, M., & Truwit, M. (2023). *Considerations for use of teaching quality measures.* EdInstruments Brief, Annenberg Institute for School Reform, Brown University. www.edinstruments.com.