Chapter 3
Evaluating Teacher Performance and Teaching Effectiveness: Conceptual and Methodological Considerations

María Paz Fernández and José Felipe Martínez

Abstract Educational theory inextricably links teachers to student learning, as the key factor mediating educational policies and student experiences in the classroom, with research consistently showing a relationship between a range of teacher and classroom variables that exert an important influence on student outcomes. This chapter highlights the key conceptual and methodological issues involved in the evaluation of teaching and teachers, with particular focus on the distinction between the concepts of *performance* and *effectiveness*. It considers the implications of assumptions and choices around *why* the evaluation is conducted, *what* is evaluated, and *how* it is evaluated, presenting a range of methods to collect data on performance and effectiveness. Additionally, we analyze issues related to the reliability and validity of resulting inferences about teacher performance or effectiveness and the implications for policy and practice. Finally, the distinctions and commonalities in evaluating performance and effectiveness in practice are exemplified through the presentation of different models of teacher evaluation.

3.1 Introduction

Teaching and learning are the central constructs of the educational process. Educational theory inextricably links teachers to student learning, as the key factor mediating educational policies and student experiences in the classroom. Educational research supports this notion empirically, consistently showing a relationship between a range of teacher and classroom variables that exert an important influence on student outcomes (e.g., Baker et al., 2010; Brophy & Goode, 1986; Darling-Hammond, 2000; Kane et al., 2013; Rivkin et al., 2005; Tucker & Stronge, 2005).

M. P. Fernández (✉) · J. F. Martínez
UCLA, Los Ángeles, CA, USA
e-mail: mpfernandez@g.ucla.edu

J. F. Martínez
e-mail: jfmtz@g.ucla.edu

J. Manzi et al. (eds.), *Teacher Evaluation Around the World*,
Teacher Education, Learning Innovation and Accountability,
https://doi.org/10.1007/978-3-031-13639-9_3

The general assumption that an improvement in teaching will lead to an improvement in learning (Goldhaber & Anthony, 2007; Hallinger et al., 2014; Hattie, 2009) underlies most teacher evaluation systems—which increasingly are been used as a policy mechanism to trigger and guide efforts to improve teacher practices and consequently student outcomes. The earliest teacher evaluation efforts initially centered on accrediting teacher qualifications, with a focus on knowledge, credentials, experience, and personal characteristics (Martínez Rizo, 2015). The underlying assumption (and often the explicit claim) was that recruiting more talented individuals or improving the qualifications of those already in the workforce would lead to better educational outcomes for students (Porter et al., 2001). As a result, most educational systems nowadays require teachers to obtain some kind of formal teaching credential or certification and/or demonstrate basic knowledge of the content they will teach. However, mounting evidence shows that static indicators of teacher *qualifications* or experience do not sufficiently explain the large variations in student achievement observed in many countries across the world (Harris & Sass, 2009). This has led to a more recent policy shift toward assessing in more detail teacher practices, or more generally the work teachers do inside and outside the classroom, and the impact these practices have on students' learning and other outcomes.

Importantly, while teaching practices and student learning outcomes are closely linked conceptually and empirically, they are also clearly distinct constructs. However, these terms are not always defined or used consistently in the literature or in educational policy. Public reports often explicitly or inadvertently conflate concepts like teacher qualifications, teacher practice, instructional quality, educational experiences, or opportunity to learn and further combine them with outcomes like student test scores, learning trajectories, etc. The resulting *constructs* are often vaguely defined and inconsistently used and may not provide a robust foundation for developing assessment instrument procedures and associated improvement processes. A sample of the literature exemplifies this conceptual inconsistency; different systems may equate student test score gains with teacher (or teaching) *impact* (Rothstein, 2016), *success* (Corcoran, 2010), *growth* (Ehlert et al., 2014), *quality* (Sass, 2008) (Hanushek & Rivkin, 2010), *performance* (Guarino et al., 2012), or *effectiveness* (Glazerman et al., 2010). Numerous researchers have warned about the dangers of reifying the empirical link between teaching and learning, arguing that student test scores cannot capture many key aspects of the broader construct of interest (Baker et al., 2010; Darling-Hammond, 2015). Teacher evaluation from this perspective is a complex undertaking that must consider in appropriate context basic qualifications, experience, and knowledge on one hand, but also the contents taught, the interactions with students around this content, and other aspects of the work of teachers beyond the classroom comprised in a rich definition of the construct *teaching* (e.g., non-academic support, administrative duties, relationships with parents, professional development, mediation with administrators, and so forth).

This chapter highlights the key conceptual and methodological issues involved in the evaluation of teaching and teachers, with particular focus on the distinction between the concepts of *performance* (the work of teachers, broadly defined) and *effectiveness* (the impact teachers can have on relevant student outcomes). The content will focus mainly on the experience in the United States because this is the country in which the debate related to these concepts has taken place the most. We consider the implications of assumptions and choices around *why* the evaluation is conducted, *what* is evaluated, and *how* it is evaluated. More specifically, summative or formative uses and purposes for the evaluation; key constructs, frameworks, and standards underlying the evaluation; and technical properties of methods and measures used to evaluate the key constructs. Ultimately, our focus is on the reliability and validity of resulting inferences about teacher performance or effectiveness and implications for policy and practice (Baker et al., 2010; Kane & Staiger, 2012; National Research Council, 2010).

3.1.1 Purposes and Consequences

The first question posed above (*why* to evaluate teachers) distinguishes among two distinct but related and often complementary uses or purposes: One formative aimed at helping teachers improve their practice by providing feedback on their performance. A second, summative purpose evaluates the teacher over some period of teaching with the goal of making decisions about the teacher (e.g., salary, tenure, dismissal, etc.). Notably, while these purposes require different processes and methods, in reality most evaluation systems seek to balance both formative and summative purposes and structures at least in paper (Bell & Kane, this volume; Wise et al., 1985).

The consequences or stakes of teacher evaluation can vary considerably across different systems. On one end, a purely *formative* system may seek to identify areas of teaching strength and weakness and offers teachers appropriate assistance, professional development, and resources to improve their practice. As accountability models imported from the private sector have made its way into education, high-stakes teacher evaluation has adopted practices such as systems of rewards or sanctions tied to improvements in student learning, including performance pay or salary adjustments, or termination of teachers who do not meet a certain criterion (e.g., Goodman & Turner, 2013; Hanushek et al., 1999; Yuan et al., 2013).

3.2 Constructs, Standards, and Frameworks

Educational accountability systems are shifting focus away from models that center on static markers of teacher qualification, to more closely assessing teacher practices or on-the-job performance (Goe, 2007). This focus requires a robust definition and operationalization of the key constructs involved and their components. As noted earlier, everyday usage of concepts like quality, competence, practice, performance, success, or effectiveness often belies their conceptual richness and the distinctions among them. It can incorrectly suggest that they are exchangeable, or at least that they have widely accepted, consistent definitions on the literature. Historically, accountability reform efforts in education focused on a set of student learning constructs as the key outcomes to improve. Standards like the common core (CC) or next generation science standards (NGSS) operationalize these constructs in terms of key contents and ideas that students should learn in each grade and may further specify what learning of varying levels of depth *looks like*. More recently, many teacher evaluation and accountability systems focus on (or additionally comprise) a set of teacher performance constructs, which are in turn operationalized and scaled into teaching standards and frameworks—these may or may not closely align to the learning standards above.

Teaching and teacher performance are complex multidimensional constructs, comprising a variety of types of knowledge, skill, attitudes, and dispositions (Darling-Hammond, 2006; Kennedy, 2008; Muijs, 2006). In a highly influential paper, Shulman (1987) enumerated the different categories of teacher knowledge, including knowledge of content, curriculum, and pedagogy, but also pedagogical content knowledge (PCK), knowledge of learners, knowledge of educational contexts, and knowledge of educational ends, purposes, and values. This multidimensional definition makes clear that teachers are always expected to know not only the content they teach, but also the most appropriate pedagogical practices and their students' needs and context. Similarly, Bransford et al. (2005) outlined three key constructs: knowledge of learners and how they learn; conceptions of curriculum content and goals; and understanding of teaching "in light of the content and learners to be taught" (p. 10). The authors emphasize that teaching, like other professions, has a social calling and a corpus of academic knowledge that has identified "systematic and principled aspects of effective teaching," supported by verifiable evidence, but also aspects related to tradition, precedent, and experience (p. 12). Reynolds (1992) outlined the competencies, understandings, and personality characteristics expected from teachers to complete the tasks of teaching. For example, the teacher should know the individual students' abilities in order to engage them effectively and also have patience with students who have trouble understanding the subject matter.

Analyzing the historical evolution of the assessment of teacher knowledge in the United States, Gitomer and Zisk (2015) identified four models of increasing proximity to teacher practice, each with different underlying premises, and associated approaches to assessment. The American Council on Education's (ACE) development of the National Teacher Examination (NTE) in the 1940s represents the first model: teachers as educated professionals, which posits merely that teachers should possess a minimum level of *intelligence, culture*, and professional preparation. The second (teacher as a content knowledge professional) grew out of concepts in cognitive psychology which "emphasized the importance of domain-specific knowledge in the acquisition of skill" (p. 38) along with concerns about how the US educational system was preparing students to compete in a globalized economy, as captured by the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983). The third model (teacher as content knowledge for teaching professional) comprises Shulman's original PCK (Shulman, 1987) and its later adaptation into *content knowledge for teaching* (CKT) (Ball et al., 2008). The most recent model conceptualizes teaching as a knowledge-rich professional practice, and teachers as "learning specialists," with primary emphasis on the application of situated knowledge to inform classroom practice (see also Guerriero, 2018).

Importantly, some authors are skeptical that "the knowledge base for teacher education is developed enough to embody in explicit standards for practice" (Stecher & Kirby, 2004, p. 6). Nevertheless, while there are no mandatory teaching standards at the national level in the United States, standards have been developed for use in many different contexts. The National Board for Professional Teaching Standards were originally published in 1989 as a "guiding framework for every teacher's development of their practice" (National Board for Professional Teaching Standards, 2016, p. 42). Teachers who voluntarily choose to become Board-Certified are expected to demonstrate that their practice meets five *core propositions*.[1] Similarly, the Interstate Teacher Assessment and Support Consortium (InTASC) developed a set of ten Core Teaching Standards outlining what "teachers should know and be able to do to ensure every PK-12 student reaches the goal of being ready to enter college or the workforce in today's world" (Council of Chief State School Officers, 2013, p. 3).[2] InTASC standards outline expected performance, essential knowledge, and critical dispositions for teachers, which have been adopted in many US states in a number of contexts, and most recently and prominently the edTPA (Educative Teacher Performance Assessment) used in dozens of states for initial teacher certification (California Commission on Teacher Credentialing, 2009; Sato, 2014).

The Danielson Framework for Teaching (Danielson, 2013) comprises four domains and 22 components of teaching. While not explicitly presented as standards, the framework operationalizes these components of teaching in terms of expected competencies and behaviors along a developmental continuum (from unsatisfactory to distinguished). The FFT has influenced or been adapted into teaching frameworks and standards that are the basis for teacher evaluation in a great number of districts in the United States and in other countries around the world. For example, the FFT is the basis for teacher evaluation systems such as the ones used in Chile (see Sun in this same volume), Peru (see Espinoza & Miranda in this same volume), New York City, and Quebec, Canada (OECD, 2013).[3]

The Australian Professional Standards for Teachers established in 2011 (revised in 2018) define seven standards, grouped into three domains of teaching (professional knowledge, professional practice, and professional engagement), which outline the expected capabilities at four stages of the teaching career (Australian Institute for Teaching & School Leadership, 2018). The Australian teaching standards also outline

[1] Subject knowledge; commitment to student learning; monitoring and managing student learning; reflecting around and learning about their own practice; and membership in learning communities.

[2] Learner development; learning differences; learning environments; content knowledge; application of content; assessment, planning for instruction; instructional strategies; professional learning and ethical practice; and leadership and collaboration.

[3] In 2020, guidelines for remote teaching were issued for the FFT, which focus on components that are thought to be most relevant for online learning and remote instruction (The Danielson Group, 2020).

the expected competencies for each level of the teaching career, associated with the educator's experience and mastery of the profession. Teachers begin as Graduate after they completed their initial training and can then progress to Proficient when they show they have achieved the seven standards. The next two levels (Highly Accomplished and Lead) are experienced teachers who work collaboratively and can be examples for others in the profession (Australian Institute for Teaching & School Leadership, 2018).

In contrast to Australia, the Teachers' Standards in England are not associated to specific stages of the teaching career and apply to almost all educators regardless of their experience. These Standards, which came into effect in 2012, are divided into two parts and outline the behaviors teachers should exhibit. Part 1 refers to teaching, stating that teachers should "act with honesty and integrity; have strong subject knowledge (…); forge positive professional relationships; and work with parents in the best interests of their pupils" (Department for Education, England, 2013, p. 10). Part 2 outlines the behaviors and attitudes related to teacher's personal and professional conduct, expecting them to "demonstrate consistently high standards of personal and professional conduct" (Department for Education, England, 2013, p. 14).

In contrast to the general frameworks and standards presented above, others are subject-specific and meant to be applied to a particular content area. A range of examples exist, including the Ambitious Science Teaching framework (Windschitl et al., 2018) and the mathematical quality of instruction (MQI) framework (Hill, et al., 2008; Hill et al., 2012; for more examples see, e.g., Bell et al., 2020; Connecticut State Department of Education, 2010; Kloser, 2014; Maine Department of Education, 2012; National Council of Teachers in Mathematics, 2000).

While most subject-specific frameworks focus on math or science, some examples may be found in the language disciplines, for example, the PLATO framework (Protocol for Language Arts Teaching Observation) for effective literacy instruction in English (Grossman et al., 2013). Frameworks can also refer to specific age groups and grades, like the Children's Learning Opportunities in Early Childhood and Elementary Classrooms (CLASS) framework (Hamre & Pianta, 2007).[4] Finally, while frameworks created in the United States and western countries typically focus on classroom behaviors and technical aspects of pedagogy, other international frameworks aim more broadly at *teacher* characteristics, competencies, and even professional and personal profiles (see, e.g., the Singapore Teaching Competency Model, which emphasizes teachers' identity as professionals charged with goals like *nurturing the whole child*, *winning hearts and minds,* or *acting in the student's interest*; Martinez et al., 2016a, 2016b).

[4] The area of emotional support encompasses the dimensions of classroom climate, teacher sensitivity, and regard for student perspectives, while classroom organization includes behavior management, productivity, and instructional learning format. Finally, instructional support is operationalized into concept development, quality of feedback, and language modeling.

3.3 Evaluating Teacher Performance and Teaching Effectiveness

Teacher *performance* evaluation or assessment aims to monitor and judge aspects of instruction and broader professional practice deemed essential or important by a system or key stakeholders. The evaluation entails collecting evidence of classroom instructional practices conducive to student learning, and others seen as important for the daily work of teachers (e.g., collaborating with colleagues or school leadership, engaging with parents and the community, etc.; Goe et al., 2008). It seeks to use approaches and methods that reflect the complexity of teaching—and more generally, teacher on-the-job performance. Authentic, contextualized information and evidence contribute to the real and perceived validity of an evaluation system and can help improve adoption and lessen distrust and resistance (Hamilton, 2005). This is also critical in cases where the evaluation is intended to support formative or improvement goals and for helping teacher education programs promote key skills and practices in teacher candidates (Darling-Hammond, 2008).

By contrast, evaluation of teaching *effectiveness* typically shifts the focus from *inputs* (e.g., teacher qualifications) and *processes* (e.g., teaching) to specific *outcomes* (e.g., student learning as captured by their scores on a standardized test (Meyer, 1996). *Effectiveness* is consequently defined as the extent of change or improvement on student learning outcomes that can be attributed to the teacher or "a teacher's ability to produce higher than expected gains in student test scores" (Goe et al., 2008, p. 5). Standardized tests are relatively easier to collect and less expensive to implement than other outcome and process measures (Cohen, 1995), and to proponents they promise consistent and *valid* comparisons across students and teachers (Papay, 2012). Advances in technology and statistics have made it easier to collect, connect, and analyze longitudinal data in new ways, particularly to create classroom- or teacher-level aggregates reflected changes in student achievement. The highest profile example of this type of approach was the Measures of Effective Teaching (MET) study, and systems of teacher evaluation inspired by it (Bill & Melinda Gates Foundation, 2010), which explicitly define *effective* teachers as those whose students exhibit more *growth* in standardized test scores (and less frequently other types of outcomes). The resulting evaluation systems are summative in spirit and rely on incentive theory, assigning monetary or other rewards and penalties for high and low effectiveness teachers, respectively (Cohen, 1995). Notably, successful teaching here is reflective of individual traits and effort, rather than "a set of learned professional competencies acquired over the course of a career" (Elmore, 1996, p. 16); while other approaches and methods are sometimes used to assess teaching (e.g., observation protocols, student surveys), these indicators are considered useful or valid mainly or exclusively insofar as they are predictive of student achievement outcomes or growth (Kane et al., 2013).

3.4 Methods and Instruments to Assess Teaching Performance and Teaching Effectiveness

3.4.1 General Considerations: Validity and Reliability

Early in the twentieth century, researchers had already identified the challenges and issues related to the *scientific* study of teaching, including selecting among many potential teaching-related constructs, occasions or instances of these constructs in classrooms, and approaches or methods to collect evidence of these constructs, each with particular strengths and weaknesses (Muijs, 2006). The key considerations from a measurement perspective are validity (the degree to which the evidence collected supports a particular inference, interpretation, or use, see AERA, APA & NCME, 2014) and reliability (the extent to which an instrument produces consistent measures of a construct across replications of a measurement procedure). The process of validation entails collecting evidence to support a proposed interpretation or use, with different kinds of evidence typically needed to support different interpretations and uses (AERA, APA, & NCME, 2014; Kane, 2006). Similarly, investigation of reliability in the case of measures of instruction ideally entails assessing the extent of measurement error from a variety of sources (e.g., raters or observers, occasions of measurement, tasks, and dimensions). The role of occasion error is particularly prominent in this context, as instruction is expected to fluctuate across contents, units, days, or even parts of lessons or days—both according to plan, and for unexpected reasons.

Validity and reliability requirements are also tightly linked to the consequences of the evaluation. High-stake evaluations (usually associated with summative purposes) may have more stringent methodological requirements, to ensure that the data used to make the decisions is adequately measuring teachers' practices. This generates additional concerns especially in the case of large-scale teacher evaluation systems (with large numbers of teachers), as the demands for methodological rigor need to be weighed against practical constraints (e.g., feasibility and cost).

A variety of methods and instruments have been developed to measure constructs related to teaching performance and effectiveness as evidence for evaluating teachers (Goe et al., 2008). Each method has different characteristics and properties and combinations of strengths and weaknesses in relation to validity, reliability, and feasibility for particular purposes and in a particular context. The following section provides an overview of a cross section of the most widely used methods and sources of evidence used to measure teaching performance and effectiveness.

3.4.2 Measures of Teaching Performance

Supervisor ratings. Information on teacher practice can be collected through ratings from individuals who supervise teachers, which can include school administrator or personnel from local or national educational agencies, researchers, or outside evaluators. These evaluations are the most common component of teacher evaluation systems in the United States, with evidence collected using a variety of specific approaches (e.g., formal or informal observations; interviews; document review), more or less structured and systematic depending on the goals and stakes (Stodolsky, 1990). The stakes can vary widely within and between systems, from formative uses focused on providing information to teachers on how to improve their practice to higher stakes uses that include decisions related to hiring or promotion (Goe et al., 2008). In general, principals are assumed to have enough contextualized knowledge about teaching performance, and studies have shown adequate reliability and positive correlation between principal ratings of teachers and student achievement (see e.g., Harris et al., 2014; Medley & Coker, 1987). At the same time, questions have been raised about subjectivity, leniency (Hamilton, 2005), reliability (Weisberg et al., 2009), and formative value, since supervisors often lack necessary substantive knowledge, particularly in higher grades (Goldstein & Noguera, 2006).

Peer evaluations. Peer ratings are attractive in teacher evaluation, because colleagues have extensive first-hand information of the knowledge and expertise required in classroom instruction and also the challenges and limitations teachers commonly face. Peer evaluation models such as *Peer Assistance and Review* (PAR)[5] rely on experienced coaches distinguished for their excellence in teaching and mentoring to provide full-time support to incoming and struggling veteran teachers. Some studies have shown that school districts that have implemented PAR have had positive results on retaining novice teachers and dismissing underperforming ones (Goldstein & Noguera, 2006; Johnson & Fiarman, 2012).

However, some evidence indicates that the benefits of peer evaluation accrue only when the evaluated and evaluator have "equivalent in assignment, training, experience, perspective and information about the setting for the practice under review" (Peterson, 1995, p. 100), which constrains the range of application and potentially its feasibility. Other potential issues with peer review include resistance to give negative evaluations to peer teachers, especially colleagues in the same school. Furthermore, these evaluations may lack the necessary credibility within teachers if there is no clear evidence of the evaluator's expertise, leading to no changes in teacher practice (Johnson & Fiarman, 2012).

[5] In these models, teachers who have been identified for their excellence in teaching and mentoring are chosen as coaches to provide support to new teachers as well as experienced colleagues who may require help. Coaches are also responsible for the teachers' formal personnel evaluations. Typically, coaches do not work in a single school, but are matched with teachers from different schools according to grade level or subject area.

Classroom observation. Observations are the most commonly used instruments for teacher evaluation and development all over the world (Bell et al., 2019; Gitomer & Zisk, 2015). In the most basic sense, this approach involves the systematic observation of live or pre-recorded lessons, during which a rater uses an observation protocol, rubric, or rating instrument to systematically register and/or assess teacher practice along a certain continuum or set of categories. Observation enjoys high face validity and has historically been seen as the Gold Standard for measuring instruction, providing direct evidence of teaching as it happens in classrooms, which can best help identify areas for improvement and professional development (Pianta & Hamre, 2009).

Classroom observation instruments can be classified as requiring low or high inference or level of subjective judgment from the rater about the teaching practices they are observing. Low inference refers to actions that observers can readily observe and record, reporting their volume or frequency (e.g., number of times students raise their hand or that the teacher asks question to all students). High-inference measures, on the other hand, require observers to assess instructional practice in terms of various *qualities* or dimensions related to specific constructs (e.g., the teacher asks high-order questions to students) (Wragg, 1999). Most widely known and used observation instruments are high-inference measures (e.g., CLASS, TALIS Video, FFT), each of which defines and operationalizes a set of distinct but related constructs of classroom practice and a continuum of quality to assess them (Martinez & Fernandez, 2019). Subject general observation instruments include the FFT (Danielson, 2013) and CLASS (Pianta et al., 2007), while examples of subject-specific instruments are the ones used in the video study of TALIS in math (OECD, 2020), PLATO in English (Grossman et al., 2013), and RTOP in science (Sawada et al., 2002).

Systematic study of observation measures in the context of teacher evaluation is far from conclusive (Martinez et al., 2016a, 2016b). In general, high-inference observation tends to show lower reliability (Muijs, 2006) or require observers to receive more intensive and expensive training to achieve appropriate levels of reliability (Bill & Melinda Gates). As was mentioned earlier, instruction can vary considerably over time, and therefore, reliability improves when teachers are observed on several occasions (albeit at increased cost). Nevertheless, observation measures have generally lower reliability and precision than traditional self-report and other standardized instruments that do not involve human judgment. Even with rigorous training and certification, high levels of reliability require several raters and occasions (Bill & Melinda Gates). Recent studies highlight these challenges faced in using even the best-known observation rubrics to support inferences and decisions involving individual teachers (Kane et al., 2011). Additional concerns relate to the potential effects the observer may have on the teacher being observed and whether the observed occasion is a representation of the teacher's typical practice or is best conceived as a high watermark (Muijs, 2006).

Teacher surveys and logs. Teacher self-reports of their practice inside or outside the classroom can range from a simple checklist of easily observable behaviors to sets of questions aimed at measuring more qualitative multidimensional constructs. Surveys can be used to study and monitor a wide range of teaching practices at scale and also to assess teachers' dispositions, attitudes, and self-efficacy, in addition to encouraging teachers' self-reflection on their practice (Goe et al., 2008). Surveys comprise a number of items (most of them close-ended) intended to measure one or several aspect or constructs of instruction and teacher practice. An advantage of teacher surveys is their cost-efficiency compared to other instruments (e.g., classroom observation), as they allow to collect data on large numbers of teachers at a relatively low cost and burden to educators. According to Mullens (1995), large-scale surveys are most useful for monitoring four areas of teacher practice: general pedagogy, professional development, instructional materials and technology, and topical coverage within courses. An example of a large-scale survey of teacher practice is the OECD Teaching and Learning International Survey (TALIS), which in 2018 included a sample of 260,000 teachers across 48 countries and economies (OECD, 2019). The Trends in International Mathematics and Science Study (TIMSS), aimed at measuring students' achievement, also collected information on teachers' beliefs and practices in 64 countries in 2019 (Mullis et al., 2020).

Disadvantages of surveys include memory error and social desirability bias, whereby teachers' responses do not reflect real practices or beliefs if they believe these would make them appear in a negative light. These disadvantages can be especially problematic when surveys are used in high-stakes teacher evaluation, such as the self-evaluation Chilean case (see Sun in this same volume). There is also evidence that teacher responses in questionnaires may not match well their instructional practice as recorded by more *authentic* measures based on classroom observations (Muijs, 2006). There are also concerns that teachers may interpret the concepts and aspects of practice in the survey different than researchers and from each other (Ball & Rowan, 2004; Mullens, 1995). For example, survey answers from two teachers may indicate they "always emphasize higher-order skills" (a 5 in a 5-point scale); but these responses may mask substantial differences across teachers, which may over- or underreport the actual frequencies (intentionally or by mistake), or have different interpretations of what is meant by *always*, *emphasize,* or *higher order*.

Teacher logs are brief surveys that are administered frequently in some cycle or period to keep a frequent and detailed record of a small number of typically narrower aspects of practice (Rowan & Correnti, 2009). Because teachers report on their practices frequently, logs reduce problems with memory and recall error prevalent present in end of year and other surveys that cover longer spans of time, resulting in better reliability and generalizability—compared to classroom observation logs which typically comprise much broader samples of occasions and offer better coverage and representation of actual practice (Ball & Rowan, 2004; Rowan & Correnti, 2009). Daily reporting of practice also tends to lessen concerns about interpretation, aggregation, and social desirability in teacher reports. Nevertheless, the advantages of specificity and frequency come at the cost of more nuanced representation of interactions between teachers and students; some researchers argue logs are only suited to studying the *amount* of content taught as opposed to *how* content was taught (Matsumura et al., 2008).

Artifacts and portfolios. Artifacts and portfolios have been used extensively in teacher induction and certification (Martinez et al., 2012). Teachers compile and typically annotate or contextualize a collection of materials and artifacts meant to illustrate their work inside the classroom. Examples of classroom artifacts include lesson plans, assignments, samples of student work, readings, and quizzes among many others. While these instruments traditionally relied on physical materials and paper copies, the incorporation of technology into the data collection process in recent years has enabled electronic portfolios that can comprise images, audio, and videos and can be managed through mobile devices (Kloser et al., 2021).

Portfolios are commonly assumed to represent the teacher's exemplary work and not necessarily their everyday instruction (Goe et al., 2008), but can be structured for daily or routine collection and monitoring typical practice and trajectories of instruction (Martinez et al., 2012). Advantages of artifacts and portfolios lie on their coverage (compared to teacher surveys) and cost (typically lower than observations), as well as strong face validity among teachers and educators, who believe these are an authentic reflection of key aspects of instruction grounded on tangible materials, and present an adequate picture of their instructional practice (Goe et al., 2008). Portfolios can be used to assess important aspects of teaching practice with reliability comparable to observations and other measures that involve human judgment (Stecher et al., 2005). Moreover, portfolio collection requires a strong cognitive commitment from teachers, which makes them valuable learning tools that encourage reflection on instructional practice (Shulman, 1998). Several large-scale teacher evaluation systems over the world are making use of portfolios as part of the instruments to gather information on teacher practices. Prominent examples of portfolios in the United States include the NBPTS certification of excellence, which requires teachers to present a comprehensive structured collection of classroom artifacts and reflections covering lessons and units across a span of months of instruction. At the other end of the teaching career path, the edTPA portfolio (Pecheone et al., 2013) is used in dozens of US states for initial teacher certification. Portfolios are also the basis for the National Teacher Evaluation System in Chile (Taut & Sun, 2014).

Disadvantages of portfolios include, on one hand, their inherent limitation in directly reflecting interactive and verbal classroom instruction and, on the other, the very substantial resources they require to develop, administer, collect, and review. Additionally, portfolios can present a considerable burden on teachers who are responsible for gathering the data over time—although when the process is framed within a professional development cycle, this *burden* is instead seen as the bulk of the work conducive to cognitive growth and learning. Finally, recent critiques of the edTPA call into question whether the psychometric properties of portfolio ratings sufficiently reflect the extent of error and thus uncertainty involved in inferences about individual teachers (Gitomer et al., 2019).

Student questionnaires. Students can be seen as one of the main sources of information on what happens inside the classroom, as they are the ones who spend more time in contact with teachers and their instruction throughout their schooling experience. Student surveys can be used to provide feedback to teachers about how their students perceive their practice, to inform school administrators and communities about average teacher practices in the school, to evaluate individual teachers, and to guide professional development (Bill & Melinda Gates Foundation, 2012a; Kuhfeld, 2017). They are increasingly used as a source of information of and for teacher practice and present important advantages over other instruments. Students' scores report individual student experiences more accurately, but aggregated at the classroom level can offer reliable composites of teacher practices that are more strongly related to student achievement than composites obtained from teacher surveys (Ferguson, 2012). Studies have shown that student surveys can be as reliable as teacher surveys (Martinez, 2012) and classroom observations (Bill & Melinda Gates Foundation, 2012a, 2012b).

Nonetheless, student surveys face significant measurement questions related to error in interpretation (especially with younger students), within- and between-level invariance, and treatment of consensus in student reports. These complexities can lead to misleading or unwarranted inferences and limit the value of the information for informing teacher learning (Schweig, 2016). The issue of within-classroom consistency deserves especial attention and is straightforward to illustrate: consider two classrooms with a mean report of 3 on a 5-point scale reflecting the challenge of assessments and quizzes. One classroom could include two groups of students with radically different perceptions: half the students not challenged (1) and the other half rather overwhelmed (5). In the second classroom, there is perfect consensus and all students reported moderate challenge (3). While both teachers receive a report that shows the same average score, this hides different patterns in responses that show students' experiences with their instruction are qualitatively very different. Appropriately reflecting within-classroom variation can thus be crucial for appropriately interpreting student survey data, and noticing differentiated or individualized instruction, or different student experiences or perspectives within the classroom.

Additional concerns focus on young children's ability to report accurately and biases (negative and positive) or inattention with older students. More broadly, students are able to report on their experiences, but are not technically qualified to assess teachers on specific areas of teaching such as curriculum and content knowledge (Goe et al., 2008). Finally, the exact wording of items can affect student responses, as items with different references can have different psychometric properties (e.g., an item worded as "my teacher asks me to read out loud" may not necessarily be interpreted in the same way as "our teacher asks us to read out loud" (Cole et al., 2011).

3.4.3 Measures of Teaching Effectiveness

Student Growth Models with Test Scores. Student achievement measures are commonly used in the United States to assess schools and teachers' effects on students' learning—they have been a staple of school reform in recent decades.

In contrast to performance measures, evaluating teachers based on student achievement places the emphasis on instructional ends (student learning), rather than means (Popham, 1971), so these models seek to determine the growth in students' achievement and attribute this to the school or the teacher. Models based on student's achievement growth effectively assume, first, that student achievement is a more direct indicator of learning than measures of teacher practice and, second, that achievement measures can accurately and validly predict success in higher education, future earnings, and aggregate economic outcomes (Hanushek & Rivkin, 2010). They thus posit that the ultimate evidence of effectiveness lies on the teacher's ability to have an effect on student learning. Indeed, early proponents argued that there was no clear evidence that teacher behavior was a good predictor of student learning, thus calling into question whether performance measures were ever appropriate to assess student learning (Millman, 1981).

Teacher evaluation based on student's achievement scores in standardized tests has been heavily criticized by experts and the broad educational community. It is argued that privileging summative over formative goals teacher evaluation approaches based on student test scores fail to offer detailed evidence necessary to guide teacher reflection and learning, which is ostensibly a fundamental necessary condition for a system that seeks instructional improvement (Amrein-Beardsley, 2008). Test-based accountability more broadly reduces the idea of "good teaching" to improvement on test scores, effectively assuming that all relevant teaching and learning information can be collected through a standardized test (Apple, 2007).

A practical concern with these instruments is their limited reach. Most standardized tests in use today measure content related to mathematics, reading, or, to a lesser extent, science. The focus on mathematics and reading (English) in the United States can be attributed to requirements from NCLB and ESSA that mandated states to test students in these subjects annually in grades 3 through 8 and then once in high school (Every Student Succeeds Act, 2015; U.S. Department of Education, 2001). Estimates suggest that the longitudinal test scores needed to produce student growth measure estimates are simply not available for as many as 50 to 60 percent of teachers across the US—a sobering reminder of the feasibility of this type of approach, even in the USA, the country that relies most extensively on standardized tests across levels of the educational system. Finally, given the high-stakes nature of these tests and the potential consequences for teachers and schools, the incentive is to reduce the hours spent on teaching subjects that will not be assessed through these tests. Evidence shows that this shift is even more pronounced in school districts serving mostly low-income and minority students, which are more at risk of sanctions for their low scores (Baker et al., 2010). Additional issues are associated with validity (e.g., whether the test measures traits that can be influenced by instruction, if the instrument is used for its intended purpose, among others) and instructional sensitivity of the tests themselves (e.g., the instrument's ability to distinguish between strong and weak instruction; Popham, 2007).

Value-Added Models (VAM). The most prominent effort to advance evaluation of teaching effectiveness in the last two decades has been the advent of so-called *Value-Added* models, which rely on students' scores in standardized tests to estimate the individual effects of a teacher on student learning growth by residualizing average students' test score gains, allowing for more precise indicators of effectiveness (Glazerman et al., 2011). The trend of using standardized tests for school assessment increased in the 1980s, with a surge in test scores used for accountability purposes toward the early 1990s (Linn, 2000). This was heightened with the passing of the No Child Left Behind Act in 2001, that stressed accountability and improvement by making schools prove their effectiveness through Adequate Yearly Progress (AYP) reports[6] (U.S. Department of Education, 2001). Since VAM are longitudinal, they can measure students' progress over time while controlling for "all of the factors that contribute to growth in student achievement, including student family, and neighborhood characteristics," isolating the effect of teachers and schools (Meyer, 1996, p. 200; Goe et al., 2008).

Along with the surge of these methods for teacher evaluation has come strong criticism from educational experts, warning both about psychometric limitations, and broader consequences of strong reliance on test scores. In the specific case of VAM, their face validity is questioned, as teachers do not understand the complex underlying statistics and cannot derive useful information for reflecting on their practice (Grossman et al., 2013). The strongest assumptions behind these models are that students' test scores are a product of their teachers' practices (i.e., a causal relationship between instruction and achievement) (Baker, et al., 2010) and that the aggregates computed can in fact reflect a causal effect. In fact, because students are not randomly assigned to teachers, the presence of bias from unmeasured variables affecting the estimates is always a strong possibility (Rubin et al., 2004). Very few studies have been able to conduct random assignment of students to teacher to establish causality with inconclusive results (e.g., Kane & Staiger, 2008; Kane et al., 2013). Baker et al. (2010) raise further concerns about the inadequacy of statistical controls to account for the student's context and the imprecision and instability of the estimates over time, class, and models (see also Darling-Hammond et al., 2012). Estimates are also inconsistent across achievement measures (Lockwood et al., 2007) which would suggest that effectiveness differs for different skills, in which case estimates should be broken down by subscore.

Student Growth Percentiles (SGP). Other teacher evaluation systems employ student growth percentiles to determine teachers' effectiveness, providing a context for a student's current achievement by locating their most recent score in a conditional distribution that depends on their prior achievement scores. In order to use this information for teacher evaluation, the students' percentiles are aggregated, and the teacher's effectiveness is determined against a defined quantity of adequate student growth whose adequacy can be determined through probabilistic (a fixed growth percentile threshold) or growth-to standard methods (the growth percentile necessary

[6] AYPs were defined as a specific amount of yearly progress in standardized test scores a school, district, or state was expected to make in a year.

to reach the desired performance level threshold Betebenner, 2009, 2011; Walsh & Isenberg, 2013).

An important feature of SGPs is that they are based on a normative conceptualization of student growth, in which the student's learning is measured in comparison with their peers, as opposed to the absolute criterion employed in VAM where the amount of growth is represented by a change in scale score points. Therefore, an advantage of SGPs over VAM is that they tend to be more accessible to teachers and school administrators and can be more easily interpreted. Although both growth models rely on complex estimations to determine the student's actual growth, SGPs provide a percentile rank that has intuitive meaning for the public (e.g., an SGP of 78 means that the student demonstrated more growth than 78% of their peers). However, this normative criterion can also be considered a limitation of SGPs, as these measures by themselves do not provide information on whether the student's relative ranking and their growth are determined to be adequate in their particular educational context (Doss, 2019).

Another perceived strength of SGPs is that they "sidestep many of the thorny questions of causal attribution", focusing on descriptions of student growth that can inform discussions about educational quality (Betebenner, 2009, p. 43). Contrary to VAM, SGPs do not require a vertical scale for the pre- and post-tests (both tests do not have to be on the same scale), so the basic requirement is that they measure the same construct. This is believed to be a more realistic constraint, as a vertical scale is a requisite to estimate VAM estimates (Betebenner, 2011).

However, SGPs present other important limitations. When compared with VAM, SGPs are more sensitive to classroom composition, as they typically do not adjust for student characteristics other than prior achievement (e.g., income, special education status, gender, etc.). This explains in part why SGPs do not perform as well as VAM when students are not randomly assigned to teachers, an assumption that tends to hold in real-life educational situations, implying that teachers who have more disadvantaged students in their class will obtain lower SGP scores than other educators (Doss, 2019; Guarino et al., 2015). Furthermore, research has shown that VAM and SPG models provide dissimilar estimates of student growth and, consequently, of teacher effectiveness, since the estimation methods are different (Goldhaber et al., 2014; Kurtz, 2018).

Student Learning Objectives (SLOs). These measures of student growth are defined as a set of goals that measure teachers' progress in achieving a certain student growth target. They differ from other measures of student growth in that they do not rely on students' scores on standardized tests, but are based on learning targets defined by teachers or educator teams. The development of SLOs follows several steps, where the teacher or education team review of standards identifies core concepts and student needs, sets goals for students, monitors student progress, and finally examines outcome data to determine next steps. Teachers are required to collect baseline and trend data from students in order to determine if they are meeting the goals set for the class. Teachers then must gather baseline and follow-up data, which can come from district assessments, student work sample, and units tests, among other sources (Lachlan-Haché et al., 2012a, 2012b).

SLOs are believed to have several advantages over other types of teacher effectiveness assessments. On one hand, they promote reflections around student results and progress, reinforcing good teaching practices, recognizing teachers' expertise, and empowering teachers as participants in their own evaluation process. On another, SLOs can be adapted to different educational contexts, allowing teacher evaluation to adjust to changes in curriculum or assessments. SLOs can also cover any subject and are not bound by the availability of standardized test scores, which tend to be limited to a few areas of knowledge (reading, mathematics, and science; (Lachlan-Haché et al., 2012a, 2012b).

However, SLOs also present several downsides. Although many states require SLOs to be "rigorous and comparable," providing guidance on acceptable measures to evaluate whether the objectives were reached, meeting the requirements of high-quality assessments and comparability across classrooms, schools, and districts, has proven challenging. Additionally, SLOs should ensure that the growth targets are ambitious while remaining attainable, avoiding the pitfall of setting goals that may be too easy to attain and that may not improve students' learning (Lachlan-Haché et al., 2012a, 2012b).

Other student outcomes. Student achievement is not the only outcome used to assess teaching effectiveness, as there is a growing consensus on the importance of non-cognitive measures that capture the range of effects of schools and teachers on students (Goe et al., 2008; Jackson, 2016; Schweig et al., 2018; West, 2016). Non-cognitive outcomes include higher-order skills like social-emotional learning, student communication and collaboration competencies, critical thinking, creativity, interpersonal competencies, and self-management, among a range of others. Recent research suggests that teachers can have a significant impact on on-time grade progression, absences, suspensions, and other proxies for non-cognitive skills (Jackson, 2016). This study also found that teachers whose practice contributes to the improvement of students' behavior are also able to improve longer-run outcomes like SAT-taking or future GPA scores.

Research is also showing teacher effects on *social-emotional learning (SEL) outcomes,* related to "knowledge, skills, and attitudes to develop healthy identities, manage emotions and achieve personal and collective goals, feel and show empathy for others, establish and maintain supportive relationships, and make responsible and caring decisions" (CASEL, 2020, p. 1). The CASEL framework encompasses five areas of SEL competence: self-awareness, self-management, social awareness, relationship skills, and responsible decision-making. To enhance students' social and emotional skills and attitudes, teachers can employ different practices in a developmentally, contextually, and culturally responsive ways, such as cooperative and project-based learning (CASEL, 2020). An example of the use of SEL as a measure of teaching effectiveness is found in Meyer et al. (2019), who use VAM to estimate the magnitude of classroom-level impacts on students' growth in SEL. The study looks at the effects of the four different constructs measured in the CORE Districts (growth mindset, self-efficacy, self-management, and social awareness), assessing the correlation between the SEL measure and achievement scores. The findings indicate that teachers who improve students' academic test performance may not be the same teachers who promote students' SEL, as there is a low correlation between classroom-level growth in SEL and classroom-level growth in ELA or math, although the growth mindset construct showed a moderately strong relationship.

Even though experts agree that non-cognitive outcomes are relevant and can legitimately be used to assess teachers, what we know about them "is extremely limited because the research has not yielded any truly informative information about how we can achieve any outcomes that we want students to learn in school other than achievement" (Good, 2014, p. 31). Good also points to the lack of consensus around the most relevant non-cognitive outcome and the cost and burden of collecting these alternative outcomes.

Other teacher measures. A range of indicators can be used to capture other relevant behaviors, dispositions, and practices of teaching more broadly defined. Examples of teacher behaviors may include simple markers like attendance, recordkeeping, participation in professional development, ethical behavior, professional interactions with the school community, and collaboration with colleagues, among others. Many systems historically relied on these types of indicators as the primary mechanism for assessing teachers, and these original evaluation systems are still in wide operation around the world as the basic infrastructure of teacher evaluation. An example of this is the teacher evaluation system currently used in the Los Angeles Unified School District in the United States, by incorporating *additional professional responsibilities* as one of the standards in their teaching framework. Within this framework, teachers are expected to maintain accurate records (e.g., track students' progress toward identified learning outcomes, manage non-instructional records, submit the records on time); communicate with families (e.g., inform about the instructional program and the student); and demonstrate professionalism (e.g., show ethical conduct, advocate for students; LAUSD, 2021a, 2021b).

3.5 Designs and Systems

In 2019, twenty-two states in the United States required teachers to be evaluated annually, a decrease from 27 states that evaluated teachers annually in 2015 (NCTQ, 2019). Classroom observations are the most common teacher evaluation measure, currently mandated in 36 states (e.g., Florida, Massachusetts, and New Mexico) and optional in another five (e.g., Arizona, Illinois, and Texas). The most widely used teacher observation protocols are the Danielson Framework for Teaching (Danielson, 2013) used in 18 states and the Marzano Causal Teacher Evaluation Model used in 11 states (Marzano & Toth, 2013). Six other states use rubrics developed either locally or externally in alignment to state standards (Close et al., 2020). Similarly, 31 states currently use student surveys for teacher evaluation, but only seven require these measures (e.g., Hawaii, Iowa, and Mississippi). Student surveys are not used for teacher evaluation in twenty states, and only New York currently prohibits their use (NCTQ, 2019). Finally, 34 states require indicators of learning growth based on student standardized test scores as part of their teacher evaluation system, up from only 15 in 2009, but down from the peak of 43 states in 2015. Of the states that require learning growth data, eight allow using other measures such as district assessments, student portfolios, or student learning objectives, instead of the state's standardized test. When it comes to the particular choice of growth model, 15 states use Value-Added models for summative evaluation, while three more report using these types of VAM scores only for formative purposes—e.g., North Carolina discontinued use of VAM scores for high-stakes personnel decisions and instead uses them to drive teacher professional development (Close et al., 2020). Finally, ten states leave the decision to use VAM scores to local education authorities—for example, in Maine, each school district can measure student growth using one of the two models: VAM or SLO indicators. In Texas, districts can select among VAM, SLOs, portfolios, or other measures to assess student growth (Close et al., 2020).

Table 3.1 presents a cross section of notable US and international teacher evaluation systems and summarizes some of their key characteristics. While not representative in any statistical or qualitative sense, the table reflects the great diversity of systems in terms of purposes, contexts, and technical characteristics and their similarities and differences—for more details about each system, refer to the links in the table.

Some systems focus mainly or exclusively on teacher performance, while deemphasizing or excluding effectiveness, either by design or in practice. For example, the Los Angeles Unified School District (the second largest in the United States) developed a Teaching and Learning Framework based on Danielson's (Danielson, 2013) and aligned to the California Standards for the Teaching Profession (LAUSD, 2021a, 2021b). Performance is assessed through classroom observations, teaching artifacts, student surveys, measures like attendance, and participation in professional development, while student test scores are used only as a benchmark for teachers to establish their own performance objectives. The Chilean Teacher Evaluation System is also based on the Danielson Framework (*Marco para la Buena Enseñanza* (MBE); Ministry of Education, Chile, 2008), but organizes evidence of performance in a portfolio comprising classroom artifacts and scores in an observation rubric (from a videotaped lesson), along with peer evaluation, supervisor ratings, and a self-evaluation rubric.

Conversely, in some systems, effectiveness is the central construct of teacher evaluation. For example, in the IMPACT system implemented at the District of Columbia Public Schools, Value-Added scores (IVA) make up 35% of a teacher's overall evaluation, while an additional 15% is assigned to a student growth measure based on SLOs. Similarly, the state of Florida classifies teachers in four levels of performance, but assigns at least 50% of the weight to VAM indicators of teacher effectiveness (S.B. 736, Student Success Act, 2010). Importantly, because student scores are only available for teachers in certain grades and subjects, schools in DC and Florida must rely on alternative assessments for large proportions of teachers—a

Table 3.1 Comparisons across teacher evaluation models

Education system	Measures	Emphasis	Framework	Combination	Purpose	Additional information
Chile	– Classroom observation – Teaching artifacts – Peer assessment – Supervisor ratings	Performance	MBE (Based on FFT)	Weighted (theoretical/policy)	Formative and summative	https://www.docentemas.cl/ (in Spanish)
District of Columbia	– Classroom observation – Student surveys – Supervisor ratings – Growth models (SLOs and/or VAM)	Effectiveness	DCPS essential practices	Weighted (theoretical/policy)	Formative and summative	https://dcps.dc.gov/page/impact-dcps-evaluation-and-feedback-system-school-based-personnel
Florida	– VAM – Classroom observation	Effectiveness	FFT, Marzano, other approved by state authority	Weighted (theoretical/policy)	Formative and summative	https://www.fldoe.org/teaching/performance-evaluation/
Los Angeles	– Classroom observation – Student survey – Student test scores – Teaching artifacts – Other measures	Performance	LAUSD Standards for Teaching (FFT + CA Standards)	Weighted (Locally determined)	Formative and summative	https://achieve.lausd.net/cms/lib08/CA01000043/Centricity/Domain/433/TLF%20Booklet.pdf

(continued)

Table 3.1 (continued)

Education system	Measures	Emphasis	Framework	Combination	Purpose	Additional information
Met Project	– Classroom observation – Growth models (VAM) – Student surveys – Teacher surveys	Hybrid	FFT, CLASS, PLATO, MQI, and UTOP	Weighted (empirical)	Formative (research-oriented)	https://files.eric.ed.gov/fulltext/ED540960.pdf
New York City	– Classroom observation – Growth models (SLOs and/or VAM)	Hybrid	FFT	Conjunctive	Formative and Summative	https://www.uft.org/sites/default/files/attachments/2020-21_Advance_FAQs_FINAL_051721.pdf

reminder of a fundamental data challenge facing systems that center on effectiveness and student test scores (Baker et al., 2010).[7] After lawsuits challenged this practice, the Florida courts explicitly determined that districts can use school aggregates to evaluate individual teachers (Paige, 2020). Both the DC and Florida systems assign the remaining 50% of the weight in the evaluation using observation measures and other indicators of performance, which individual districts are able to select from approved lists.

The systems in Florida and New York City Schools (the largest district in the United States) exemplify the common hybridization or conflation of the two central concepts underlying this chapter, performance and effectiveness. In New York, eight indicators derived from classroom observations are used for summative *performance* assessment, while the remaining fourteen are used exclusively for non-evaluative feedback. Interestingly, the number of observations each year is determined by the teacher's previous ratings—fewer observations for highly *effective* teachers and more for *ineffective* teachers (New York City Department of Education, 2019). While New York also evaluates teachers using measures of student learning, the model de-emphasizes individual accountability based on effectiveness. A committee with administrators and union members identifies measures, target populations (e.g., different subgroups of students at the classroom, grade, or school level), and even the model (e.g., VAM or goal setting around SLOs).

As for the approach for combining measures, a common hybrid approach combines weighting and conjunctive/disjunctive decision rules or tables. For example, in NY, a teacher rated *ineffective* in the performance measure, and *highly effective* in the measure of student learning is overall classified as *developing*. States like Colorado, Louisiana, or Pennsylvania have implemented similar decision tables. Among systems that use compensatory models, theoretical or policy weights are the commonly used (e.g., DCPS, Florida, Chile) but a variety of other approaches exist. A prominent example is the LAUSD system which frees school sites to determine how to combine information across measures (Los Angeles Unified School District, 2019). The Measures of Effective Teaching study, while not an operating system per se, deserves special mention here as the largest ever to measure teacher performance and effectiveness in thousands of classrooms using multiple observation protocols including FFT (Danielson, 2013), CLASS (Hamre & Pianta, 2007), PLATO (Grossman et al., 2013), MQI (Hill et al., 2008), and UTeach (UTOP; Walkington & Marder, 2018), teacher and student surveys (Ferguson, 2012), supervisor ratings, and even a test of pedagogical content knowledge. Researchers assessed how predictive each measure was of teacher value-added estimates based on standardized test scores and tests of higher-order conceptual understanding (Bill & Melinda Gates Foundation, 2010). The study was very influential in the US during the 2000s and 2010s among other things because it is one of few to randomly assign students to teachers to yield clearer causal effects. However, the various measures were found to

[7] Schools can adopt commercially available tests or develop their own, provided these are "rigorous, aligned to content standards, and appropriate for the teacher's classes and students" (District of Columbia Public Schools, 2011, p. 2; Gitomer & Joyce, 2015).

correlate only weakly and inconsistently to VAM scores, and the authors ultimately emphasized the importance of balancing the weights assigned to performance and effectiveness indicators for high-stakes teacher evaluation—effectively signing away the explicit emphasis on empirical weights that was originally at the core of the study design.

Notably, the results of teacher evaluation conducted over the last few years under this great variety of designs and systems are converging in classifying a great majority of teachers in the highest levels of performance. For example, in Florida, 98% of teachers statewide are rated highly effective or effective, with only 0.6% classified as developing and 0.1% unsatisfactory (Florida Department of Education, 2018), and similar proportions are commonly observed elsewhere (see, e.g., Anderson, 2013; Dynarski, 2016; NCTQ, 2017).

Finally, it is important to note that, as is commonplace across the US and internationally, all the systems in the table claim both summative and formative goals and uses of the measures collected. In Chile, for example, teachers classified as basic or unsatisfactory must complete professional development courses and engage in self-reflection and collaborative peer work to address weaknesses identified in the evaluation, but can eventually face dismissal if they continue to underperform (Taut & Sun, 2014). The DC IMPACT system similarly combines summative consequences for teachers (incentives and potentially dismissals) with individual formative feedback on four areas: instructional practice, student achievement, instructional culture, and collaboration.

3.5.1 Combining Measures to Evaluate Teaching Performance and Teaching Effectiveness

The discussion above makes it apparent that multiple instruments and methods are necessary to provide sufficient information to evaluate teacher performance and effectiveness. Indeed, multiple measures provide a more comprehensive image of both performance and effectiveness (Goe & Croft, 2009), as each of the instruments and measures described earlier is well suited to capture some performance or effectiveness constructs (in some context), but limited or ill-suited to capture others. In addition to improved construct coverage, research shows that multiple measures can produce more stable or precise categories to classify teachers (De Pascale, 2012; Steele et al., 2010), limit score inflation (NCTQ, 2015), and reduce incentives for gaming the system (Steele et al., 2010) among others. Perhaps even more importantly, evidence from multiple measures is needed to provide rich, usable feedback to teachers and thus is essential for constructing strong systems of professional development parallel to the evaluation (Baker et al., 2010; Duncan, 2012). This can also help increase of buy-in among stakeholders (Glazerman, et al., 2011) and identify and reduce adverse impact in time (De Corte et al., 2007).

There are three main approaches to combining evidence from different instruments and constructs (Martinez et al., 2016a, 2016b). *Conjunctive* models assess each measure separately and summarize the information using a joint decision rule—e.g., teachers meet the standard if they obtain a rating of *basic* or above in the observation measure and rank in the top 8 deciles in the student survey and student learning outcomes. This reduces false *positives/passes* by requiring adequate performance in each construct or component (e.g., performance and effectiveness). Conversely, *disjunctive* or complementary models require meeting a criteria for only some measures—e.g., score of *basic* or above in at least two of three measures. This reduces false *negatives/fails* and is preferred when some dimensions are more important than others. Finally, *compensatory* models create a single linear composite index synthesizing the information in the measures—this weighted average allowing high performance on one measure to compensate for lower performance on another (Brookhart, 2009). Weights can be set empirically (e.g., factor analysis, regression coefficients) or theoretically (e.g., through stakeholder negotiation).

Each of these models has advantages and drawbacks and can be used to maximize specific properties of the resulting joint inferences (Mihaly et al., 2013). Importantly, they can also lead to different classifications and decisions for individual teachers (Martinez et al., 2016a, 2016b). In this context, Martinez et al., (2016a, 2016b) suggest that balanced theoretical or policy weights have important advantages because they not only offer desirable psychometric properties in terms of composite reliability and consistency over time, but more importantly reflect a broad stakeholder consensus about the importance of different aspects of teaching performance and teacher effectiveness—a potential powerful hortatory instrument for policy adoption and implementation.

3.6 Conclusions and Implications

Educational improvement efforts centered on teacher evaluation are typically conceptualized around two related but distinct targets of assessment: teacher performance or teaching effectiveness. From the discussion presented above, it is apparent that these approaches rely first on a series of assumptions about the nature and components of *teaching*, a very complex multidimensional construct that is often defined inconsistently by educators, researchers, policymakers, and the public. In addition, these efforts and resulting systems involve assumptions and choices around conceptual and methodological aspects involved in assessing this target construct, and also the most impactful policy mechanisms for exerting influence on it, and the people and organizations involved. For example, Kane & Bell (in this same volume) discuss critical points of distinction between teacher evaluation systems conceived primarily for summative or formative goals.

Importantly, many of these considerations go beyond the strictly technical and relate to broader societal and institutional goals and contexts at the national or subnational level—where a broad range of social and political priorities, pressures, and stakeholders typically play a defining role in spearheading, shaping, modifying, and in some cases ending teacher evaluation systems (see Zorrilla & Martinez in this same volume).

In this chapter, we tried to highlight the complexities associated with these assumptions and choices, the subsequent systematic collection of information and evidence to assess what teachers do (teacher performance), and the effect teachers have on specific student outcomes (teaching effectiveness). The former concept relies on models and frameworks that outline the ideal competencies, practices, and attitudes of teachers. The latter focuses on measuring and improving outcomes, and attaching incentives to the evaluation, with the expectation that this will affect instruction. While effectiveness is often linked with summative goals, and performance with formative objectives, the more useful distinction is at the level of individual instruments or measures, which may be more conducive to formative or summative uses. For example, classroom observation protocols tend to be used in formative teacher evaluation, as they are a source of direct evidence of teaching as it happens in classrooms, which can be used to identify areas of improvement and professional learning for teachers. Conversely, Value-Added Models (or similarly, student growth percentiles) are seen as more summative in nature, as they focus on teachers' ability to improve student outcomes and do not directly offer evidence to guide professional learning or improvement. Importantly, most teacher evaluation systems in operation would reject the summative label; even those with a very strong focus on estimating teacher *effectiveness* typically claim (either explicitly or implicitly) to also have formative value or serve formative goals.

To serve these dual objectives, systems typically rely on the use of multiple measures. While the notion that teacher evaluation requires multiple measures is nearly universal, this idea, like teaching, belies great conceptual and methodological complexity. On one hand, as our chapter outlined, instruments and measures have distinct strengths and weaknesses and may be advantageous for different purposes and in different contexts, inevitably presenting substantive, technical, and practical tradeoffs to developers of teacher evaluation systems. Moreover, different ways of combining information derived from these measures rest on different assumptions and can have direct implications for the inferences made about teaching and teachers. Because no approach to combining measures consistently outperforms the others on strictly technical grounds, systems should thus explore the approach that most closely aligns with their goals and that allows to best illuminate the relevant aspects of the *teaching* construct. Perhaps most importantly, the idea of *combining* the measures into a single final score for each teacher implies a loss of information that in principle would seem counter to the more formative goals these systems typically espouse, as information about specific aspects of the multidimensional construct teaching best illuminated by each instrument is blended into a single ostensibly unidimensional measure (Martinez et al., 2016a, 2016b). Instead, systems should aim to *make combined use* of the information provided by multiple measures, to best utilize the full extent and detail of information provided by each one for formative or summative purposes, or both.

There is mounting evidence, including much reflected in other chapters in this volume, that irrespective of whether performance or effectiveness is the main *narrative* focus, the technical rigor of the instruments is not sufficient to sustain teacher evaluation systems—which additionally require thoughtful implementation, explicit and meaningful focus on improving teacher practice or performance on the ground, and realistic consideration of the institutional, policy, and political context. Without these elements in place, there are no psychometric or statistical techniques, either existing or future, that will enable education systems to sustainably and productively evaluate teachers in very large volumes, on a tremendously multidimensional construct, in complex contexts, and for high-stakes purposes.

References

AERA, APA, NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65–75.

Anderson, J. (2013, March 30). Curious grade for teachers: Nearly all pass. *New York Times.*

Apple, M. W. (2007). Ideological success, educational failure? On the politics of no child left behind. *Journal of Teacher Education, 58*(2), 108–116.

Australian Institute for Teaching and School Leadership. (2018). *Australian Professional Standards for Teachers.* AITSL.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Shepard, L. A., et al. (2010). Problems with the use of student test scores to evaluate teachers. *EPI Briefing Paper* (278).

Ball, D. L., & Rowan, B. (2004). Introduction: Measuring instruction. *The Elementary School Journal, 5*(1), 3–10.

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*(5), 389–407.

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.

Bell, C. A., Klieme, E., & Praetorius, A.-K. (2020). Conceptualising teaching quality into six domains for the Study. In OECD, *global teaching insights technical report* (pp. 1–24). OECD Publishing.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories.* The National Center for the Improvement of Educational Assessment.

Bill & Melinda Gates Foundation. (2010). *Learning about teaching: Initial findings from the measures of effective teaching project.* Bill & Melinda Gates Foundation.

Bill & Melinda Gates Foundation. (2012a). *Gathering feedback for teaching. Research Paper.* Bill & Melinda Gates Foundation.

Bill & Melinda Gates Foundation. (2012b). *Asking students about teaching. Policy and practice brief.*

Bransford, J., Darling-Hammond, L., & LePage, P. (2005). Introduction. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 1–39). Jossey-Bass.

Brennan, R. L. (2001). Some problems, pitfalls, and paradoxes in educational measurement. *Educational Measurement: Issues and Practice, 20*(4), 6–18.

Brookhart, S. M. (2009). The many meanings of multiple measures. *Education Leadership, 67*(3), 6–12.

Brophy, J., & Goode, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). MacMillan.

California Commission on Teacher Credentialing. (2009). *California standards for the teaching profession (CSTP).*

CASEL. (2020). *CASEL'S SEL framework: What are the core competence areas and where are they promoted?* CASEL.

Close, K., Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of state's teacher evaluation systems post–every student succeeds act. *Education Policy Analysis Archives, 28*(58), 1–26.

Cohen, D. K. (1995). Rewarding teachers for student performance. In S. Fuhrman, & J. O'Day (Eds.), *Rewards and reforms: Creating educational incentives that work.* Jossey-Bass.

Cole, M. S., Bedeian, A. G., Hirschfeld, R. R., & Vogel, B. (2011). Dispersion-composition models in multilevel research: A data-analytic framework. *Organizational Research Methods, 14*(4), 718–734.

Connecticut State Department of Education. (2010). *Common core of teaching: Foundational skills.* CSDE.

Corcoran, S. P. (2010). *Can teachers be evaluated by their students' test scores? Should they be? The use of value-added measures of teacher effectiveness in policy and practice.* Annenberg Institute for School Reform.

Council of Chief State School Officers. (2013). *InTASC model core teaching standards and learning progressions for teachers 1.0.* CCSO.

Danielson, C. (2013). *The framework for teaching evaluation instrument, 2013 edition.* Danielson group.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–44.

Darling-Hammond, L. (2006). Constructing 21st-century teacher education. *Journal of Teacher Education, 57*(3), 300–314.

Darling-Hammond, L. (2008). Reshaping teaching policy, preparation, and practice: Influences of the national board for professional teaching standards. In R. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the national board for professional teaching standards (Advances in Program Evaluation)* (Vol. 11, pp. 25–53). Emerald Group Publishing Limited.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132–137.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal tradeoffs between selection quality and adverse impact. *Journal of Applied Psychology, 92*, 1380–1393.

De Pascale, C. (2012). Managing multiple measures. *Principal, 91*(5), 6–10.

Department for Education, England. (2013). *Teachers' standards: Guidance for school leaders, school staff and governing bodies.* DFE.

District of Columbia Public Schools. (2011). *Teacher-assessed student achievement data (TAS) guidance.* DCPS.

Doss, C. J. (2019). *Student growth percentiles 101: Using relative ranks in student test scores to help measure teaching effectiveness.* RAND Corporation.

Duncan, A. (2012, agosto 22). *Change is hard*. Retrieved from US Department of Education: https://www.ed.gov/news/speeches/change-hard

Dynarski, M. (2016). *Teacher observations have been a waste of time and money.* Brookings Institution.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. J. (2014). The sensitivity of value-added estimates to specification adjustments: Evidence from school- and teacher-level models in missouri. *Statistics and Public Policy, 1*(1), 19–27.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–26.

Every Student Succeeds Act, Title I Section 1111(2)(B)(III)(vi) (2015).

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan, 94*(3), 24–28.

Florida Department of Education. (2018). *2017–18 District educator evaluation ratings.* Retrieved from Archived Statewide District Evaluation Results: http://www.fldoe.org/teaching/performance-evaluation/archive.stml

Gitomer, D. H., & Joyce, J. (2015). *A review of the DC IMPACT teacher evaluation system.* National Research Council.

Gitomer, D. H., & Zisk, R. C. (2015). Knowing what teachers know. *Review of Research in Education, 39*, 1–53.

Gitomer, D. H., Martinez, J. F., Battey, D., & Hyland, N. E. (2019). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal, 58*(1), 3–31.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating evaluation systems.* Brown Center on Education Policy at Brookings.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Brookings Institution.

Goe, L. (2007). *The link between teacher quality and student outcomes: A research synthesis.* National Comprehensive Center for Teacher Quality.

Goe, L., & Croft, A. (2009). *Methods of evaluating teacher effectiveness.* National Comprehensive Center for Teacher Quality.

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

Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *The Review of Economics and Statistics, 89*(1), 134–150.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the model matter? Exploring the relationship between different student achievement-based teacher assessments. *Statistics and Public Policy, 1*(1), 28–39.

Goldstein, J., & Noguera, P. A. (2006). A thoughtful approach to teacher evaluation. *Educational Leadership, 63*(6), 31–37.

Good, T. L. (2014). What do we know about how teachers influence student performance on standardized tests: And why do we know so little about other student outcomes? *Teachers College Record, 116*, 1–41.

Goodman, S. F., & Turner, L. J. (2013). The design of teacher incentive pay and educational outcomes: Evidence from the New York City bonus program. *Journal of Labor Economics, 31*(2), 409–420.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education, 119*, 445–470.

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2012). *Can value-added measures of teacher performance be trusted?* Education Policy Center at Michigan State University.

Guarino, C. M., Reckase, M. D., Stacy, B., & Wooldridge, J. M. (2015). A comparison of student growth percentile and value-added models of teacher performance. *Statistics and Public Policy, 2*(1), 1–11.

Guerriero, S. (2018). *Teachers' pedagogical knowledge and the teaching profession: Background report and project objectives.* OECD Publishing.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educational Assessment, Evaluation and Accountability, 26*(1), 5–28.

Hamilton, L. (2005). Lessons from performance measurement in education. In R. Klitgaard & P. C. Light (Eds.), *High-performance government* (pp. 381–405). RAND Corporation.

Hamre, B. K., & Pianta, R. C. (2007). Learning opportunities in preschool and early elementary classrooms. In R. C. Pianta, M. J. Cox, & K. L. Snow (Eds.), *School readiness & the transition to kindergarten in the era of accountability* (pp. 49–84). Paul H. Brookes Publishing Co.

Hanushek, E. A., & Rivkin, S. G. (2010). *Using value-added measures of teacher quality.* CALDER - Urban Institute.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1999). *Do higher salaries buy better teachers?* NBER Working Paper No. 7082.

Harris, D. N., & Sass, T. R. (2009). *What makes for a good teacher and who can tell?* CALDER working paper.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: a comparative analysis of teacher effectiveness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73–112.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hill, H. C., Blunk, M., Charalambous, C., Lewis, J., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction, 26*, 430–511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56–64.

Jackson, C. K. (2016). *What do test scores miss? The importance of teacher effects on non-test score outcomes.* NBER.

Johnson, S. M., & Fiarman, S. E. (2012). The potential of peer review. *Educational Leadership, 70*(3), 20–25.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation.* NBER Working Paper 14607.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching.* Bill & Melinda Gates Foundation. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2680&filename=MET_Gathering_Feedback_Research_Paper1.pdf

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011, Summer). Evaluating teacher effectiveness: Can classroom observations identify practices that raise achievement? *Education Next* (pp. 55–60).

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

Kennedy, M. M. (2008). Sorting out teacher quality. *Phi Delta Kappan, 90*(1), 59–63.

Kloser, M. (2014). Identifying a core set of science teaching practices: A Delphi expert panel approach. *Journal of Research in Science Teaching, 51*(9), 1185–1217.

Kloser, M., Edelman, A., Floyd, C., Martinez, J. F., Stecher, B., Srinivasan, J., & Lavin, E. (2021). Interrogating practice or show and tell? Using a digital portfolio to anchor a professional learning community of science teachers. *Journal of Science Teacher Education, 32*(2), 210–241.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the tripod student survey. *Educational Assessment, 22*(4), 253–274.

Kurtz, M. D. (2018). Value-added and student growth percentile models: What drives differences in estimated classroom effects? *Statistics and Public Policy, 5*(1), 1–8.

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012a). *Student learning objectives as measures of educator effectiveness: The basics.* American Institutes for Research.

Lachlan-Haché, L., Cushing, E., & Bivona, L. (2012b). *Student learning objectives: Benefits, challenges, and solutions.* American Institutes for Research.

LAUSD. (2021a, April 3). *History of EDST*. Retrieved from https://achieve.lausd.net/Page/11782#spn-content

LAUSD. (2021b). *Teaching and learning framework.* LAUSD.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V.-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47–67.

Los Angeles Unified School District. (2019). *2018–2019 EDS final evaluation report for teachers and non-classroom teachers: Administrator handbook.* LAUSD.

Maine Department of Education. (2012). *Common core teaching standards.* MDE.

Martínez Rizo, F. (2015). La evaluación del desempeño docente. Una propuesta para la educación básica en México. In G. Guevara Niebla, M. T. Melendez Irigoyen, F. E. Ramon Castaño, H. Sanchez Perez, & F. Tirado Segura (Eds.), *La evaluación docente en México* (pp. 64–95). INEE-Fondo de Cultura Económica.

Martinez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23*(3), 305–326.

Martinez, J. F., & Fernandez, M. P. (2019). Evaluación docente con indicadores múltiples: Consideraciones conceptuales y metodológicas en torno a la validez. In J. Manzi, M. R. Garcia, & S. Taut (Eds.), *Validez de Evaluaciones Educacionales en Chile y Latinoamérica* (pp. 531–562). Ediciones UC.

Martinez, J. F., Borko, H., & Stecher, B. (2012). Measuring instructional practices in middle school science using classroom artifacts. *Journal for Research in Science Teaching, 49*, 38–67.

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016a). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis, 38*(4), 738–756.

Martinez, J. F., Taut, S., & Schaaf, K. (2016b). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation, 49*, 15–29.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* ASCD.

Matsumura, L. C., Garnier, H. E., Slater, S. C., & Boston, M. D. (2008). Toward measuring instructional interactions "at-scale." *Educational Assessment, 13*, 267–300.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research, 80*(4), 242–247.

Meyer, R. H. (1996). Value-added indicators of school performance. In E. A. Hanushek & D. W. Jorgenson (Eds.), *Improving America's schools: The role of incentives* (pp. 197–223). The National Academies Press.

Meyer, R., Pier, L., Mader, J., Christian, M., Rice, A., Loeb, S., Hough, H., et al. (2019). *Can we measure classroom supports for social-emotional learning? Applying value-added models to student surveys in the CORE districts.* PACE.

Mihaly, K., McCaffrey, D., Staiger, D., & Lockwood, J. R. (2013). *A composite estimator of effective teaching (MET Project).* The RAND Corporation.

Millman, J. (1981). Student achievement as a measure of teacher competence. In *Handbook of teacher evaluation* (pp. 146–166). Sage.

Ministry of Education, Chile. (2008). *Marco para la Buena Enseñanza.* MINEDUC.

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation, 12*(1), 53–74.

Mullens, J. E. (1995). *Classroom instructional processes: A review of existing measurement approaches and their applicability for the teacher followup survey.* U.S. Department of Education.

Mullis, I. V., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science.* TIMSS & PIRLS International Study Center.

National Board for Professional Teaching Standards. (2016). *What teachers should know and be able to do* (2nd ed.). NBPTS.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* U.S. Department of Education.

National Council of Teachers in Mathematics. (2000). *Principles and standards for school mathematics.* NCTM.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy.* National Academy of Sciences.

NCTQ. (2015). *State teacher policy yearbook: National summary.* National Council on Teacher Quality (NCTQ).

NCTQ. (2017). *Running in place: How New teacher evaluations fail to live up to promises.* NCTQ.

NCTQ. (2019). *State of the states 2019: Teacher & principal evaluation policy.* National Council on Teacher Quality (NCTQ).

New York City Department of Education. (2019). *Advance guide for educators 2019–2020.* NYCDE.

OECD. (2013). *Teachers for the 21st century: Using evaluation to improve teaching.* OECD Publishing.

OECD. (2019). *TALIS 2018 results: Teachers and school leaders as lifelong learners* (Vol. 1). OECD Publishing.

OECD. (2020). *Global teaching insights: A video study of teaching.* OECD Publishing.

Paige, M. (2020). Moving forward while looking back: How can VAM lawsuits guide teacher evaluation policy in the age of ESSA? *Education Policy Analysis Archives, 28*(64), 1–18.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123–141.

Pecheone, R. L., Shear, B., Whittaker, A., & Darling-Hammond, L. (2013). *2013 edTPA field test: Summary report.* SCALE.

Peterson, K. D. (1995). *Teacher evaluation: A comprehensive guide to new directions and practices.* Corwin.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*(2), 109–119.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2007). *Classroom assessment scoring system.* Paul H. Brookes.

Popham, W. J. (1971). Performance tests of teaching proficiency: Rationale, development, and validation. *American Educational Research Journal, 8*(1), 105–117.

Popham, W. J. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, 146–155.

Porter, A., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259–297). AERA.

Reynolds, A. (1992). Getting to the core of the apple: A theoretical view of the knowledge base of teaching. *Journal of Personnel Evaluation in Education, 6*, 41–55.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rothstein, J. (2016). *Can value-added models identify teachers' impacts?* IRLE—UC Berkeley.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. *Educational Researcher, 38*(2), 120–131.

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103–116.

S.B. 736, Student Success Act. (2010). *St.* FL.

S.B. 736, Student Success Act, Section 1012.343(3)(a)1 (2010).

Sass, T. R. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy.* CALDER—Urban Institute.

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education, 65*(5), 421–434.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the agreement of student ratings. *Learning Environments Research, 19*(3), 441–462.

Schweig, J., Baker, G., Hamilton, L. S., & Stecher, B. M. (2018). *Building a repository of assessments of interpersonal, intrapersonal, and higher-order cognitive competencies.* RAND Corporation.

Shulman, L. (1998). Teacher portfolios: A theoretical activity. In N. Lyons (Ed.), *With portfolio in hand* (pp. 23–37). Teachers College Press.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

Stecher, B. M., Wood, A. C., Gilbert, M., Borko, H., Kuffner, K. L., Arnold, S. C., & Dorman, E. H. (2005). *Using classroom artifacts to measure instructional practices in middle school mathematics: A two-state field test (CSE Report 662).* CRESST.

Stecher, B., & Kirby, S. N. (2004). *Organizational improvement and accountability: Lessons for education from other sectors.* RAND Corporation.

Steele, J., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems.* The RAND Corporation.

Stodolsky, S. S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175–190). Corwin Press.

Taut, S., & Sun, Y. (2014). The development and implementation of a national, standards-based, multi-method teacher performance assessment system in Chile. *Education Policy Analysis Archives, 22*(71).

The Danielson Group. (2020). *The framework for remote teaching.* The Danielson Group.

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning.* Association for Supervision and Curriculum Development.

U.S. Department of Education. (2001). *No child left behind act (Executive Summary).* U.S. Department of Education.

Walkington, C., & Marder, M. (2018). Using the UTeach observation protocol (UTOP) to understand the quality of mathematics instruction. *ZDM Mathematics Education, 50*, 507–519.

Walsh, E., & Isenberg, E. (2013). *How does a value-added model compare to the colorado growth model?* Mathematica Policy Research.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project.

West, M. R. (2016). Should non-cognitive skills be included in school accountability systems? Preliminary evidence from California's CORE districts. *Evidence Speaks Reports, 1*(13), 1–7.

Windschitl, M., Thompson, J., & Braaten, M. (2018). *Ambitious science teaching.* Harvard Education Press.

Wise, A. E., Darling-Hammond, L., McLaughlin, M. W., & Bernstein, H. T. (1985). Teacher evaluation: A study of effective practices. *The Elementary School Journal, 86*(1), 60–121.

Wragg, E. C. (1999). *An introduction to classroom observation.* Routledge.

Yuan, K., Le, V., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., & Springer, M. G. (2013). Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. *Educational Evaluation and Policy Analysis, 35*(1), 3–22.