Improving Teacher Evaluation Systems: Making the Most of Multiple Measures

Edited By Jason A. Grissom and Peter Youngs

Reviewed by *Jan L.H. Frank*
University of St. Thomas

Accountability, standardization, assessment, evaluation—all are very familiar terms in education these days. Calls for greater accountability from schools and teachers related to student learning have led to increased use of standardized methods of collecting data to provide assessment information and ultimately to evaluate the effectiveness of teachers, programs, and schools. Few would argue against the need for accountability in any profession, and developing standardized approaches can help ensure fairness of assessment. Problems arise, however, when one confuses assessment and evaluation. Assessment is the means by which data are collected so that a description of "what is" can be developed. Frequency of specific behaviors, raw scores on teacher-created tests, and averages of all work completed with a specified period of time can all be ways to provide assessment data to detail what is happening within a classroom or for a particular student. The data are then analyzed so that one can determine what actions should be continued and what might need to be changed in order to reach goals. Evaluation, on the other hand, is a judgment made using the data that are collected and analyzed through assessment—a determination of "good or bad," "effective or ineffective." The distinction is important. When systems do not recognize the difference, final judgments (evaluations and decisions about people's positions) can be made based on limited or non-existent data/assessment. We are so apt to leap to judgment. Having a focus regarding the purpose for collecting data is essential, as is being clear about the level (teacher, classroom, school, district) at which the data analysis will be used.

Jason A. Grissom and Peter Youngs offer an edited volume that focuses on measures of teacher effectiveness and the implementation of multiple measures of these. As they note in their introductory chapter, "increased efforts to hold teachers accountable for their performance via multiple measures of their impacts on their students and schools have been among the most important education policy shifts of the last decade" (p. 1). The introductory chapter synthesizes "lessons learned" and provides guidelines for practice and policy in the area of teacher evaluation. They note that there appear to be "two main purposes of teacher evaluation: helping teachers improve their instruction and informing human capital decisions in schools and districts" (p. 169-170). They caution that clear evidence of reliability and validity of the many evaluation measures available must be established, and that both teachers and administrators must feel confident that measures provide accurate ratings of teacher performance. The remainder of their chapter provides a description of research on new evaluation approaches and guidelines for their use. Before moving into the rest of the volume, Grissom and Youngs remind the reader that if the purpose of teacher evaluation is formative, and not for high stakes decision-making, then concerns about reliability and validity are less serious. It is also important to determine at what level the data regarding teachers is being used—is it for the teachers themselves, is it for school level administrators, or is it for district decision-making?

Measures of Teacher Effectiveness

Improving Teacher Evaluation Systems provides a description of the most common measures currently used to determine teacher effectiveness as well as an analysis of some of the important challenges of using the measures appropriately. Chapters 2 through 8 are concerned with what these measures are, how multiple measures fit together, and the challenges related to collection and interpretation. Teacher observations continue to be the most common measure of teacher effectiveness, despite numerous concerns. For example, training for observers and standardization of observation approaches are both rare. The sensitivity of observation tools is also a concern—do these tools consider the context of individual classrooms, or simply assume a "one size fits all" approach? Can observations detect changes in teaching over time? Finally, one can question whether there even exists a definition of teacher effectiveness that allows for any kind of accurate measurement.

Using classroom observations as a measure of teacher effectiveness is considered by Julie Cohen and Dan Goldhaber in Chapter 2. The limitations of observational measures that are discussed include questions about stability and rater reliability, and whether such measures adequately consider contextual factors. The authors also ask whether observations are sensitive to changes in practice—for example, will an observation measure be sensitive enough to provide data indicating a change in teacher performance due to professional development? Of equal importance is determining whether a teacher's future performance is improved as a result of the information received from performance measures such as observations. In Chapter 3, Robert C. Pianta and Bridget K. Hamre continue the discussion of teacher observation measures. They contend that a theory of teacher effectiveness that delineates measureable elements must be developed. In addition, they argue for the standardization of observation approaches and effective training that would lead to the certification and recertification of observers.

Using measures other than observation also raises questions. If one relies on a narrow scope of measures, will the wide range of contributions that teachers actually make in terms of student growth and achievement and in terms of the school as a whole be considered? Grissom, Susanna Loeb, and Christopher Doss argue in Chapter 4 that the reliance on student test score growth as the only measure of teacher effectiveness or contribution is much too narrow. Their discussion revolves around the limited correlation between value-added measures and other aspects of teacher performance that principals tend to rate highly: building students' interpersonal skills, maintaining positive relationships with other faculty and staff, and making contributions to school leadership. They suggest that research should seek to understand and evaluate the more varied contributions teachers make to students and the school as a whole.

In response to these concerns, a "value-added" approach as an assessment of teacher effectiveness has been advocated. This approach uses measures that attempt to isolate the contributions of individual teachers to a student's learning by statistically adjusting for differences, like prior academic performance and socioeconomic background, among students. However, Sean P. Corcoran, in Chapter 5, questions whether a value-added approach adequately provides teachers and school leaders with "instructive, actionable information that they can use to make meaningful changes to instructional practice" (p. 59). He does advocate, though, for value-added measures to be used to provide an "early-warning signal," if paired with other

measures, to help identify teachers who demonstrate lagging performance and need professional development. Value-added measures are also useful to researchers who need to validate other assessments of teacher effectiveness or evaluate the impact of policies, programs, and interventions.

A final concern about measures used to evaluate teachers is whether the same tools can be used for all teachers, no matter their grade level, content area, or type of student. In Chapter 6, Nathan D. Jones examines whether the same tools can be used to evaluate both general educators and special educators. Existing measures may not sufficiently capture the complexity of special educators' work. In addition, the co-teaching that often occurs with general educators and special educators is not at all conducive to tools such as value-added measures—there is no way to adequately differentiate the impact of the general educator from that of the special educator in terms of student learning. Jones also notes that these concerns can also apply to general educators, who often share instruction across grade levels and who may use different effective teaching practices for different subject areas.

Additional measures to evaluate teachers are discussed in the remaining chapters of the first section of the book. The idea of including student surveys to measure teacher effectiveness is discussed by Ryan Balch in Chapter 7. He argues that student surveys have the potential to identify specific areas that teachers can focus on for improvement, can provide similar information as observations at lower cost, and have some research support as measures of teacher effectiveness. He provides a "validation framework," a way to provide a body of evidence that supports the use of student surveys, and includes methods for using student surveys effectively. Balch concludes his chapter acknowledging the concern that students may not be capable of providing accurate feedback, especially for high stakes evaluation, and suggests future research to explore such concerns. Teachers would need to accept the student feedback and be willing to implement improvements based on that feedback, as well, something that many teachers may be reluctant to do.

Finally, Chapter 8, written by Youngs and Andrea Whittaker, looks at the Education Teacher Performance Assessment (edTPA) that is now being used by a number of states to evaluate teacher candidates. They note that while there is evidence to support the validity, reliability, and fairness of the assessment, this evidence is still evolving, making it important to look for additional evidence beyond the essential psychometric properties—for example, whether we can say that the edTPA is an appropriate measure for the high stakes decisions of granting initial teacher licensure. In addition, since the edTPA was developed for teacher candidates and not experienced teachers, there is a need for examining the predictive and consequential validity of the edTPA; therefore, Youngs and Whittaker advocate that the edTPA not be used for annual evaluations of practicing teachers.

Implementing Multiple Measures

The second section of the book, Chapters 9 through 13, examines the implementation of multiple-measure teacher evaluation systems and shares the experiences of teachers and principals using the new measures, including the challenges they have faced in implementation. In Chapter 9, Min Sun, R. Brock Mutcheson, and Jihyun Kim discuss "… teachers' reports on

the use of teacher evaluation to improve their instruction and what types of school supports aid in this process" (p. 103), including follow-up professional development programs and timely feedback from observations. After sharing their study, they note some limitations in being able to interpret the relations between providing supports for teachers and improvements in teachers' instruction, and ultimately increasing student achievement as a result. The data on teachers' perceptions and use of evaluations were self-reported by each teacher, creating a common source bias. Sun, Mutcheson, and Kim recognize this in their chapter and recommend research designs to mitigate such problems and potentially aid in establishing validity of the evaluation measures.

Principals' use of teacher evaluation measures is explored by Tim Drake, Ellen Goldring, Grissom, Marisa Cannata, Christine Neumerski, Mollie Rubin, and Patrick Schuermann in Chapter 10. They found that in districts with well-developed teacher evaluation systems, the evaluation process provides leaders with a means of delivering needed teacher support and that dismissal processes occurred only after the support processes were provided. Principals rely on classroom observations as a tool to provide specific feedback to ineffective teachers. Of particular interest (and concern) is the paradox the authors note: "… the definition of ineffective includes reliance on value-added measures of student achievement, but the means to determine if a teacher is making progress toward becoming more effective rests principally with another data source, teacher observations" (p. 128). That means there exists a mismatch between the measure used to determine "ineffectiveness" and the ones used to indicate improved performance. The authors, therefore, suggest ensuring that the observation measures being used have reliability and validity so that both teachers and administrators can trust the data collected with these tools.

In Chapter 11, Morgaen L. Donaldson and Casey D. Cobb discuss implementing standards-based observations and student learning objectives in teacher evaluation. After looking at Connecticut's implementation of these practices, they conclude that both show promise for determining the effectiveness of teachers. Observations, however, are perceived as more beneficial by the teachers because such measures provide concrete, usable data that they are familiar with.…. Implementing the collection of data on students' progress towards meeting learning objectives as a measure of teacher effectiveness requires more support at the district, school, and educator levels. There is also the challenge of determining valid and appropriate student learning objectives for determining teacher effectiveness, and it is difficult to attribute any student growth to individual teachers because there are so many factors that can affect learning in complex, interdependent ways.

In Chapter 12, Gary T. Henry and J. Edward Guthrie address the concern that value-added scores are often simply reported to teachers and principals without explanations of any connection to daily practice. The authors researched North Carolina, a state that relies on a system of evaluation with multiple measures: principal ratings, student surveys, and classroom observations. These measures are then associated with the teachers' value-added scores to determine any correlations. The strength of this system is that it provides meaningful feedback to teachers as to what they can improve to increase students' learning and achievement. Simply being told to "improve a value-added score" is meaningless in terms of day-to-day practice—and this is one of the major arguments teachers have against evaluation systems. Henry and Guthrie's suggestions provide the basis for moving forward in evaluation and reaching its true purposes—improved teaching practice and increased student learning.

In the final chapter of section two, Venessa A. Keesler and Carla Howe confirm much of what was already discussed in earlier chapters as they examine the implementation of a new educator evaluation system in Michigan. This chapter provides a good overall summary for the reader, as well as an overview of the specific application of these ideas in Michigan. Recommendations they offer include ensuring the validity of instruments used to quantify teacher effectiveness, requiring and supporting an appropriate and high quality professional development system when conducting educator evaluations, and determining how evaluation results are being used to inform improved practice. Again, the clear message is that the validity of any measures used to evaluate must be assured and that professional development support is essential if evaluation is to be used to improve teacher effectiveness.

Conclusion

Grissom and Youngs, as editors, provide a concise look at teacher evaluation. The fact that they include discussions about multiple measures is particularly important. Far too often we depend on single measures, like student scores on standardized tests, to determine teacher effectiveness. The contributors for each chapter discuss the research base for each measure and share both concerns and suggestions regarding their use. The book is well-organized and would be helpful for those concerned with knowing more about teacher evaluation measures. Due to the many facets involved in teacher evaluation and the issues surrounding validity, an overall visual (a graphic, concept map, or chart) would have been useful to help the reader better conceptualize the complex nature of assessment, as well as what evidence we currently have to support the validity of measures for professionals in the field of education compared to where questions and concerns still need to be addressed.

For example, unfortunately unaddressed in this book are the impacts of factors such as race, gender, and class, and their intersectionality, on measurements and assessments for students and teachers. It is easy to simply assume that traditional means of determining validity and reliability are sufficient, but rarely do those means consider the influence of demographic data. Assessment and evaluation tools, whether they are used to determine student achievement and growth or teacher effectiveness, must be analyzed for evidence of insensitivity, exclusionary language, and assumptions about marginalized groups.

Overall, I can see this volume supplementing a research methods course for teachers since the discussions of the research are very readable and provide insights into studies that would be of particular interest to teachers. In addition, this book would provide excellent education examples for a course on measurement and assessment at either the undergraduate or graduate level.

Author Notes
Jan L. H. Frank is an Associate Professor in the Department of Teacher Education at the University of St. Thomas.

Correspondence concerning this article should be addressed to Jan L.H. Frank at jlhfrank@stthomas.edu.

Reference

Grisson, J.A., & Youngs, P. (2015). *Improving teacher evaluation systems: Making the most of multiple measures.* New York: Teachers College Press.