Improving Teacher Evaluation Systems

Making the Most of Multiple Measures

EDITED BY

Jason A. Grissom
Peter Youngs

Hawker Brownlow
EDUCATION

Contents

TCP7262 • 9781760017262 • © 2016 Hawker Brownlow Education

CHAPTER 1

Making the Most of Multiple Measures

Jason A. Grissom
Peter Youngs

Teacher evaluation has become serious business. The days of principals hastily completing a teacher evaluation checklist based on one or two short observations and marking everyone *satisfactory* are vanishing in most school districts. Indeed, 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. Spurred by the Obama administration's Race to the Top program and its requirements for state waivers from No Child Left Behind, and, in some cases, investment from the Bill and Melinda Gates Foundation and other philanthropies, many states and districts have implemented evaluation systems that combine statistical estimates of teachers' impacts on student achievement (e.g., "value-added" measures) with scores from newly designed rubric-based observation protocols, feedback from student surveys, and other metrics to produce more comprehensive measures of teacher performance than have ever been available before.

The ability to gather more, more rigorous, and more specific information about teacher performance would seem to be an unequivocally positive policy development, given the potential of this information to inform and improve all kinds of school decisions, from how to target supports for teacher development to which teachers should be hired or dismissed to how teachers are compensated. As these systems take hold, though, teachers, administrators, researchers, and policymakers are raising important questions about the measures these systems rely on, what uses of those measures are appropriate or inappropriate, and how new teacher evaluation systems are—and are not—changing teacher practice, school leader decisionmaking, and the culture of schools. In addition, multiple-measures-based teacher evaluation systems have ignited substantial controversy.

Systems to collect, manage, analyze, and report evaluation data cost schools substantial money and time. Value-added measures, which use statistical techniques to calculate teachers' impacts on student achievement growth, have been attacked as biased and misleading. New observation systems have been rolled out in many places with insufficient training for raters and too little attention to ensuring fidelity to instruments and protocols, with some educators raising concerns that they provide information no more useful than what was provided previously. The use of student surveys has been derided by many teachers worried either that students cannot answer them appropriately—a big concern for teachers of younger students—or that they will simply be used as a means for students to get back at teachers they do not like. In sum, many teachers are left feeling that their evaluations do not in fact reflect the quality of the work they do.

Despite the controversy surrounding these data-gathering and evaluation systems and the fact that the changes—positive and negative—they are creating are not well understood, we see continued expansion not only of the systems themselves, but of their use to drive personnel decisions and promote accountability in other domains, such as teacher preparation and principal evaluation. The goal of this volume, therefore, is to take stock of what we have learned about the impacts and challenges of data-intensive teacher evaluation systems from these initial years of development and implementation, and identify challenges for practitioners and researchers in the years ahead. We argue that rigorous teacher evaluation systems have the potential to promote school improvement, but only if the systems are carefully designed and implemented and the data they generate are interpreted and used appropriately. The chapters that follow, penned by scholars and policymakers working at the cutting edge of research and policy in this area, speak to what we know and what remains to be known about evaluation measures themselves, the implementation of evaluation systems, and the use of evaluation data. They also make recommendations for state policymakers and district administrators moving forward with such systems.

The book is loosely organized into two sections. The first section, comprising Chapters 2 through 8, focuses attention on the most important measures of teacher performance currently used by multiple-measures evaluation systems, what their properties are, how they fit together, and the challenges inherent in collecting and interpreting the measures well. The second section—Chapters 9 through 13—turns to the implementation of multiple-measures teacher evaluation systems and what we have learned about the experiences of teachers and principals as they use these new measures, and challenges that school, district, and state administrators face in implementing them. A final chapter pulls together lessons

learned and guidelines for practice and policy synthesized from across the two sections.

The first section begins with a chapter by Julie Cohen and Dan Goldhaber that compares the two major components of nearly every multiple-measures-based teacher evaluation system: value-added measures of teacher performance and classroom observation measures. The authors contrast what can be learned from each measure type, what conditions are necessary for the accuracy of each one, and the most important sources of error for each. The authors point out that substantial research on value-added measures has illuminated a variety of concerns about their limitations and biases and may have pushed practitioners to more highly value observation-based measures, when in fact observation instruments face many potential sources of inaccuracy and bias that have not been investigated with the same intensity.

Chapter 3 follows up on issues of accuracy and reliability of teacher observation tools. Drawing on their experiences developing and implementing the Classroom Assessment Scoring System (CLASS) for classroom observations, Robert C. Pianta and Bridget K. Hamre outline a variety of fundamental components that must be in place to facilitate high-quality observation systems at scale that can produce consistent, reliable measures of teacher practice. These include broad components such as a protocol that is grounded in a theory of effective teaching to motivate measurement, as well as features related to the use of that protocol, such as standardization of observation approaches and effective training, certification, and recertification systems for observers. The usefulness of the information gleaned from teacher observations is unlikely to be very high in the absence of such investments.

In Chapter 4, Jason A. Grissom, Susanna Loeb, and Christopher Doss examine whether value-added measures are good proxies for other facets of teachers' contributions to their schools beyond increases in standardized test scores in mathematics and reading. Specifically, they test how well value-added correlates with assessments given by principals in low-stakes interviews of different areas of teachers' job performance. The authors find that value-added tends to correlate well with principals' assessments of a teacher's ability to produce high test performance from students and develop higher-order thinking but weakly or inconsistently with other dimensions, such as building student interpersonal skills or contributing to leadership in the school, suggesting that value-added captures only a relatively narrow range of a teacher's contributions. When asked which teachers they wished to retain, principals named teachers they rated highly on these other performance dimensions but not necessarily those with high value-added scores. The authors conclude that evaluation systems that

privilege value-added measures are likely to overlook important contributions to the school that many teachers make and that the multidimensional nature of teachers' work requires a multiple-measures approach to evaluation.

The fifth chapter continues this close look at value-added measures. In this chapter, Sean P. Corcoran discusses the potential pitfalls of using teacher value-added data for teacher evaluation and other high-stakes decisions. He argues that the presumed benefits of value-added scores—their conciseness, their statistical objectivity, the fact that they show variation in teacher effectiveness when other traditional measures often identify everyone as similarly high-performing—are more than outweighed by numerous shortcomings that researchers have begun to catalogue but that policymakers have often been unaware of or discounted. These include numerous sources of bias in the scores, measurement error that produces unstable estimates, and problems with face validity, among others. As a summative measure, they also provide scant guidance to practitioners about what in a teacher's classroom works well or what areas need to be targeted for improvement. The core of Corcoran's argument is that leaders and policymakers should lower their expectations about what can be learned from individual teachers' value-added and avoid giving such measures an outsized role in evaluating job performance.

Chapter 6 focuses on evaluation for a population of teachers in an area given little attention in policy debate and research to this point: special education. Nathan D. Jones considers both value-added and teacher observation measures in the context of special education. He highlights several important features of special education teachers' work that pose particular challenges for evaluation systems geared toward more typical classroom teachers, including that their instruction is often very specific to individual students, with learning objectives distinct from those of general education classrooms. We know little about the properties of student test score–based performance measures for teachers of special education students or about the appropriateness of the most frequently used classroom observation instruments for special education teachers. Jones underscores the need for additional policy discussion and research in this important area.

The seventh chapter turns from value-added and teacher observations to focus on student surveys. Ryan Balch notes that student surveys are potentially useful in systems of teacher feedback and evaluation because they can identify specific areas of focus for teacher improvement and are less resource-intensive than teacher observations while providing potentially similar information. He goes on to synthesize the small literature base on student assessment of teachers and describe a framework for assessing and

TCP7262 • 9781760017262 • © 2016 Hawker Brownlow Education

ensuring the validity of student surveys as measures of teacher practice. He illustrates this framework with the Survey of Teacher Practice that he developed and makes available through My Student Survey, a student survey development and administration company. Balch argues that thoughtfully and carefully constructed student surveys can provide valid and reliable measures of many aspects of teachers' instruction, though student surveys also present important challenges, particularly around building and maintaining teacher buy-in, that districts must address if such measures are to realize their potential for aiding teacher improvement.

Chapter 8 concludes the initial section by reviewing research on edTPA, a teacher performance assessment that is being used widely with preservice candidates. In their chapter, Peter Youngs and Andrea Whittaker first consider ways that classroom observation instruments can be employed to evaluate content-specific instructional practices, as well as some limitations associated with these tools. Next, the authors describe edTPA and review research on edTPA and similar teacher assessments, including evidence related to reliability and validity. Finally, they discuss ways in which edTPA and observation tools represent distinct responses to pressures linked to accountability and explain how edTPA can be used in tandem with observation instruments to evaluate teaching candidates' content-specific instructional practices.

The second section, which begins with Chapter 9, focuses on evaluation system implementation and data use. In this initial chapter, Min Sun, R. Brock Mutcheson, and Jihyun Kim examine how teachers in two rural school systems report using teacher evaluation information for their own instructional improvement. They found this kind of utilization of evaluation information to be more common among early-career teachers on probationary contracts and in schools with more intensive, higher-quality professional development around data use and principals who provided more useful evaluation feedback. Their results suggest that appropriate school supports can push teachers to make use of performance information to increase the effectiveness of their practice.

Chapter 10 turns attention to principals' use of teacher evaluation measures. Timothy A. Drake, Ellen Goldring, and colleagues draw on interview data from principals in six urban school systems that have implemented multiple-measures teacher evaluation systems to examine how the data from those systems are used in teacher dismissal processes. The authors found that rhetoric raising concerns that new teacher evaluation systems are largely about targeting low-performing teachers for dismissal generally is not reflected in principals' experiences with these systems. Instead, evaluation measures for low-performing teachers are used to place teachers on growth or improvement plans to support their growth in

areas in which they are weak. Although the documentation associated with teacher growth plans also provides evidence for dismissal proceedings, school and district leaders often view dismissal as a by-product of an unsuccessful attempt to help a teacher improve rather than as a primary goal of the evaluation system.

In Chapter 11, Morgaen L. Donaldson and Casey D. Cobb investigate teachers' and school leaders' experiences with the implementation of Connecticut's new multiple-measures teacher evaluation system. In Connecticut's system, standards-based observations are coupled with student learning objectives, an alternative to value-added models for documenting student learning based on educators' goals. They find that classroom observations were typically welcomed by teachers, who wanted more observations and more feedback to assist them with their instruction, though administrators found them challenging to implement because of time and other capacity constraints. Student learning objectives posed their own capacity challenges, as educators struggled to set appropriate and rigorous learning goals, though the goal-setting process itself was motivating for many teachers. The authors also note that the growing use of student learning objectives in state evaluation systems requires much more research attention than currently is being devoted, particularly to their psychometric properties, which are poorly understood.

In Chapter 12, Gary T. Henry and J. Edward Guthrie outline how principals can incorporate teacher value-added scores into a process for teacher improvement. Drawing on value-added, teacher observation, and student survey data from North Carolina, they demonstrate how correlations between practices—measured by observations and surveys—and value-added in the same year and over time can be used to identify and prioritize specific practices to target for improvement. The system they describe can enable principals to provide evidence-based advice to teachers about the practices that are most likely to lead to improvements in student achievement in their classrooms.

Chapter 13 approaches teacher evaluation implementation issues from the perspective of senior personnel in state education agencies. Venessa A. Keesler and Carla Howe describe efforts by the Michigan Department of Education to support school districts as they enact new teacher evaluation systems. These efforts include providing technical assistance to districts related to the use of classroom observation systems, student learning objectives, and student growth data; supporting the development of interim assessments for use in obtaining information on teacher performance; and maintaining a focus on using evaluation to support teachers' professional development as well as to meet accountability demands. In addition to technical challenges associated with using value-added models, the authors

TCP7262 • 9781760017262 • © 2016 Hawker Brownlow Education

also note that a lack of coherence among state educational policies, along with limited resources in their agency, have affected their ability to support districts as they have implemented new evaluation measures.

In the final chapter, we look across the contributions to this volume and the related literature on multiple-measures teacher evaluation systems to summarize some key implications that can be drawn from this growing body of research. In particular, we conclude that while the information from such systems can be used to improve teacher practice and improve personnel decisions in schools, state and district leaders must think carefully about the design of multiple-measures systems and pay close attention to how they are implemented to realize that potential. As currently put into practice, we worry that measures employed in many school systems have reliability and validity properties that are too questionable to be used for summative evaluation and associated high-stakes decisions. Fortunately, these properties can be improved with attention and resources. Until that time, formative uses of evaluation measures to identify areas for teacher growth that can be addressed through professional development and support may be the highest-value use for much of this information.

With regard to final thoughts in introducing this book, we want to encourage policymakers, administrators, and researchers to continue exploring effective ways to support the process of teaching and the process of teacher evaluation. As the chapters in this book attest, an emphasis on accountability alone is unlikely to result in large-scale improvements in teaching and learning. Instead, schools, districts, and states need to consider how the use of multiple measures of teacher evaluation can be integrated with school-based and external professional development activities. In many schools, teachers' colleagues, instructional coaches, and principals can potentially be very helpful resources in the area of instructional improvement. But this potential will be realized only if and when policymakers place greater emphasis on ways in which teacher evaluation measures can be used for formative purposes to support teacher development.