Teacher Evaluation Systems: A Literature Review on Issues and Impact

Kathleen Williams, McNeese State University
Dustin Hebert, Louisiana Tech University

Abstract

Teacher evaluation systems are associated with teacher quality, accountability, performance observations, and support. These systems are typically comprised of multiple measures including at least observations of teaching and student performance data reflecting teacher impact. Many criticisms of these systems have emerged not necessarily from the measures themselves but from how they are used and the consistency with which the larger system is implemented. Concerns like evaluator training, reliability of results, distinctions between teacher quality and teaching quality, and repercussions of an ambiguous system for individual teachers, just to name a few, are recurring themes in the literature. In all, these systems are purported to foster teacher professional growth; however, ensuring clarity of purpose, sound accountability measures, and formative utility of results is a crucial milestone before the validity of these systems can be recognized.

Keywords: teacher evaluation systems, teacher accountability, teacher observation

Introduction

Since public schools are public institutions, the primary purpose of teacher evaluation systems is to hold educators accountable to the public, which funds their profession (Danielson, 2011). Teacher evaluation systems serve to assess teacher quality and to promote school improvement through professional development (Toch, 2008). They also yield and identify variations in observed teacher qualities in order to identify individuals whose practice could benefit from feedback and professional development opportunities (Hill & Grossman, 2013). However, teacher evaluation systems that differentiate among teachers and teacher qualities need to have valid and reliable results (Danielson, 2011; Shakman et al., 2012).

To consistently and accurately assess and ensure teacher quality through teacher evaluation systems, there must first be a shared definition of *good teaching* among all stakeholders (Danielson, 2011). An effective teacher evaluation system measuring teacher quality goes beyond a generic rubric or checklist and includes classroom observations, student and parent surveys, and student achievement scores (Hill & Grossman, 2013; Toch, 2008). The Bill and Melinda Gates Foundation (2013a) emphasized that having an accurate teacher evaluation system ultimately depends on the evaluation method's recognition of the multifaceted components of teaching.

There has been a significant emphasis on incentivizing teacher quality through teacher evaluations systems since teacher quality was identified as the most important factor affecting student achievement (Looney, 2011; Muijs et al., 2014; Papay, 2012). Dating back more than a decade, teacher evaluation systems have become a method of holding teachers accountable to their school leaders, district supervisors, and state governing bodies (Rivkin, Hanushek, & Kain, 2005). Additionally, teacher evaluation systems have become teacher accountability systems that emphasize improving standardized test scores in order to raise student achievement (Ahn, 2013). Teacher evaluation systems most commonly include the value-added models and standards-based classroom observations, which are both evaluation systems adopted to promote student achievement by focusing on teacher effectiveness (Papay, 2012).

Marshall (2005) argued that these teacher evaluation systems are inefficient and ineffective in achieving their purpose of improving teaching and student learning. This leads to systems that lack credibility with superficial and inconsistent teacher evaluations (Toch, 2008). Ultimately, in order to produce widespread, effective teacher evaluation systems, there must be efforts to understand the results of these measures so that teachers understand the systems' implications and how to impact student learning positively (Papay, 2012).

Teacher Evaluation Relevant to Teacher Quality

Danielson (2011) defined teacher quality as professional credibility that is identified by an evaluation system. Prior to 2001, teacher quality was predominately identified and measured by teacher experience, certification, and education levels. However, since then, several studies have shown few correlations between these factors and teacher effectiveness, new teacher evaluation methods have been adopted that observe and measure teacher performance (Harris et al., 2014; Hinchey, 2010; Stumbo & McWalters, 2011). These new teacher evaluation methods define teacher quality as teaching practices and characteristics that raise student achievement and performance. These evaluation methods attempt to measure teacher effectiveness based on this definition (Kupermintz, 2003).

Ultimately, it is essential that evaluators have a shared understanding of the definition of high-quality teaching and the profession's multifaceted components in order to assess teacher performance accurately (Danielson, 2011). Darling-Hammond (2012) reported that evaluators need to distinguish between teacher quality and teaching quality. Darling-Hammond further specified that teacher quality is the encompassed personal traits and skills that an individual brings to teaching. On the other hand, teaching quality refers to strong instructional practices that enable a range of students to learn. Therefore, teaching quality plays an important role in teacher quality. Looney (2011), on the other hand, argued that there is no widespread, accepted definition of teacher quality; however, Looney did specify that teacher quality can be assessed through sets of measurable standards. Nevertheless, Harris et al. (2014) warned that the choice of evaluation tool guides and affects a teacher's

demonstration of professional skills and qualities that are evaluated with the tool in question. Therefore, it is important to choose an evaluation system with consistent and clear standards that yields reliable results.

Additionally, the focus of evaluation systems should be teacher performance qualities that promote student achievement (Wayne & Young, 2003). Ultimately, teachers will become more successful in raising student achievement when evaluation systems accurately focus on teacher performance and effective characteristics (Hinchey, 2010). However, a consistent need of evaluations is that the evaluation systems must encourage effective teaching methods while retaining highly effective teachers and their practices (Darling-Hammond & Ball, 1998).

Evaluation Systems as Accountability Measures

Over the past two decades, federal legislation has incentivized states nationwide to raise student achievement through rigorous academic standards, increased student expectations, and assessment-based school accountability programs (Gordon, Kane, & Staiger, 2006; Muijs et. al, 2014; Rivkin, Hanushek, & Kain, 2005). Several states have adopted new, more rigorous curricula and evaluation methods that determine teacher effectiveness. These new teacher evaluation methods have been adopted, though, based on studies that show few correlations between teacher effectiveness and teacher experience, certification, and education levels, which was previously the baseline for determining teacher retention and effectiveness (Harris et al., 2014; Stumbo & McWalters, 2011). New teacher evaluation systems focus on student achievement scores as a determination of teacher effectiveness and as a way for holding teachers accountable to student performance standards. These summative evaluations are preferred for their quality assurance and accountability measures (Danielson & McGreal, 2000). For educators, the results of these accountability measures establish and determine teacher promotion, tenure, dismissal, and compensation (Harris et al., 2014).

As a result of the 2009 Race to the Top (RTTT) program, accountability systems use teacher evaluation methods to influence individual teachers using a short-term reward system (Ahn, 2013; Harris et al., 2014). Ahn (2013) noted that these incentives and accountability policies implemented at the school level by principals may improve performance level and efforts of existing teachers. Ahn argued further that when pay is associated with performance, schools usually see an improvement in student achievement scores. On the other hand, Harris et al. (2014) found that these accountability measures may influence who chooses to enter the teaching profession or deter some altogether from the profession. Nevertheless, high-stakes accountability measures only exacerbate the teachers' stress levels (Danielson, 2007). There is a particular concern for novice teachers who are unfamiliar with the stresses of teaching and who may feel pressure dealing with the new accountability systems (Roberson & Roberson, 2009). Like Harris et al. (2014) and Danielson (2007), others raise questions concerning the feasibility and desirability of teacher accountability systems (Sartain et al., 2011).

Sartain et al. (2011) investigated individual teacher responsibilities to annual student learning gains in response to newly implemented accountability systems. Sartain et al. identified that teaching is a collective, rather than solely individual, pursuit and that any policies involving teacher accountability as a reflection of individual student achievement or growth needs to reflect this fact. Since schools are relying more heavily on collaborative teaching, correlating one student's performance to a single teacher is becoming more difficult even though that approach is central to contemporary evaluation systems, thus questioning the accuracy of linking individual student performance scores to individual teachers as well as the equality of these systems. It is inferred that evaluators should apply accountability system's results to the school instead of individual teachers (Danielson & McGreal, 2000).

The equality of accountability systems is further questioned when evaluation systems require evaluators to make judgments on teaching practices (Danielson & McGreal, 2000). A common concern with performance evaluations is when novice teachers are compared and measured to the same level of effectiveness by evaluators as veteran teachers. In situations such as this, novice teachers struggle to adjust their teaching practices to align with expectations that may have been established based on effective practices of veteran teachers whose practices had been refined after years of successful teaching (Roberson & Roberson, 2009). Additionally, Ahn's (2013) research found that accountability systems impact educators of all levels and experiences, whether it guides their teaching methods during observations or encourages them to teach the test. These accountability measures influence the characteristics of teachers, which further impact the learning environment, student experiences, and student performance, both in the short- and long-term (Harris et al., 2014).

Teacher Observations as Evaluation Measures

Observations were a method used by administrators and supervisors to survey the classroom environment and teacher-child interactions (Reinking, 2015). Historically, teacher performance has been assessed by observation checklists with relatively little concern or association to student achievement and teacher quality (Hill & Grossman, 2013). These forms or surveys included items focused on direct and verbal forms of teaching practices from a set list (Danielson & McGreal, 2000). States have adopted improved instruments to evaluate teacher performance through observations that align with specific guidelines in the federally funded RTTT program in addition to other sources (Reinking, 2015). These new instruments, like Charlotte Danielson's Framework for Teaching, yield evaluations conducted by expert evaluators to assess teacher performance and behaviors relative to specific expectations (Stumbo & McWalters, 2011). The new standards-based teacher observations have been found to provide more instructional guidance to teachers and encourage best practices that increase student achievement (Papay, 2012; Stumbo & McWalters, 2011).

Observation tools of the past were implemented as a formative evaluation experience that required an observer to collect descriptive data on predetermined skills and characteristics of a teacher's performance in the classroom (Danielson & McGreal, 2000). Therefore, the tools must primarily have clearly defined skills and characteristics that specify levels of performance (Papay, 2012). However, Danielson and McGreal (2000) argued that the forms associated with these evaluation systems do not define the systems. It is the structure of the evaluation process and the professional conversations surrounding the observation that make an effective teacher observation evaluation system. Effective standards-based teacher observation evaluation systems must extend beyond the forms used and include three essential elements: (a) a clear definition of the domain of teaching, incorporating the standards for proficiency in teacher performance; (b) specific methods and procedures assessing aspects of teaching; and (c) trained evaluators who make consistent judgments on observed performances.

Several advantages of standards-based teacher evaluation systems have been documented over the traditional checklist classroom observations. Within these systems, new observation evaluations require the evaluator to cite clear evidence of teaching practices during the observation, allowing for a much richer view of a teacher's instructional practice (Papay, 2012). When teachers demonstrate strong teaching methods measured by classroom observations, their students tend to show higher academic growth regardless of previous performance scores and socioeconomic status (Daley & Kim, 2010; Sartain et al., 2011). In a similar study, students who learned from the most effectively rated teachers from these observation evaluations were found to outperform their peers by as much as one grade level from those who learned from the least effective teachers (Looney, 2011). Furthermore, the teaching standards on which standards-based observation evaluations are based have research-driven data, which links them with student achievement (Darling-Hammond, 2012).

However, since the implementation of the teacher observation evaluations, teachers have argued this evaluation is subjective and bias-ridden (Papay, 2012). One of the most prominent concerns among teachers is that evaluators' scores may be influenced by prejudices against the teacher, especially since many of the evaluators are immediate supervisors of the teachers being evaluated (Hill & Grossman, 2013; Papay, 2012). According to Hill and Grossman (2013), any form of inaccuracy in the observation evaluations compromises the diagnostic function of the observations. This will further hinder any opportunities to improve instructional practices and meaningful feedback that is central to teacher observation evaluation systems. Nevertheless, having highly-qualified and well-trained evaluators who have a clear and precise understanding of the standards on the observation evaluation rubric as well as a clear understanding of instructional proficiency eliminates much of the subjective bias (Papay, 2012).

Ho and Kane (2013), who conducted a study on fair and reliable observation systems administered by school personnel with the Bill and Melinda Gates Foundation in the Measures for Effective Teaching Project,

found observers rarely used the highest and lowest ratings, which identified teachers as exemplary or unsatisfactory, respectively. Most observers scored teachers in the middle of any given observation tool's rating range. Likewise, participating administrators' scores differentiated more among teachers, with administrators scoring their own personnel .1 point higher than leaders from other schools. Ho and Kane also found that an observer's first impression of a teacher tended to linger and impact other observation evaluations of that same teacher. Based on these findings, it was concluded that having more than one observer raises the reliability of the observation evaluation scores. Further, there was a 60% increase in reliability of an observation evaluation when the observation was conducted in single 15-minute instances instead of observing the full hour. Ultimately, findings suggested a district could monitor the reliability of classroom observations in order to ensure a fair and reliable system for teachers.

Conclusion

The nationwide shift to teacher accountability has led to widespread adoption of standards-based, observation-driven evaluation systems. These systems are based on teaching practices and characteristics that are associated with teacher effectiveness. During these evaluations, teachers receive scores and feedback from school leaders that highlight areas of strength and areas where performance improvement is needed.

Although these systems are purported to yield formative and constructive evaluations that foster teacher performance improvement, educators have criticized these systems for subjectivity and implementations contradictory with improving teacher performance. How the tools and the implementation processes yield accurate assessments of teacher quality remain questioned. These standards-based observation evaluations are praised for having individualized teacher feedback, which leads to reflective discussions and teacher personal growth; however, these systems are, ultimately, accountability tools. Thus, having a genuine discussion on professional growth is difficult when teachers feel this criticism or feedback comes with negative repercussions of accountability. In all, the result is often deemed punitive rather than formative and constructive.

Although the tools used to evaluate teacher performance are, mostly, rubrics with performance criteria and indicators stated, evaluators' interpretations of those criteria and assumptions or preconceptions of a lesson's quality and rigor impact the ratings awarded. These factors have led to questioning the validity and reliability of the standards-based observation results, which are imperative given the accountability outcomes for teachers.

Current standards-based evaluation tools do identify teacher strengths and areas for improvement as well as provide baselines for professional reflection, and such reflection should help improve teacher quality and highlight professional development needs. The value this evaluation process offers reinforces its effectiveness; however, the accountability component that accompanies an evaluation tool distracts from the professional

development and reflection process. Therefore, it is imperative to maintain this process with an emphasis that it is a tool for educators' professional growth and development.

Since marrying accountability with a tool intended to promote professional growth has led to divisiveness, policymakers should reevaluate the purpose and components of teacher evaluation systems. First, clarity in these systems' purpose is needed. Are these systems used to judge performance, foster professional growth, or both? Second, if the intention of these systems is to hold educators accountable, a more well-rounded system comprised of evaluation and accountability components that reflect the multifaceted components of effective teaching is warranted. This would address concerns of fairness, accuracy, and credibility. Finally, to demonstrate that results are used formatively rather than punitively, systems should include reflect the use accountability results through fair evaluation measures to foster teacher quality through professional growth that emphasizes both teaching quality and teacher quality. If the ultimate goal is to have successful teachers and students in all classrooms, teacher evaluation systems must demonstrate not only the capacity to measure success but also to support it.

References

Ahn, T. (2013). The missing link: Estimating the impact of incentives of teacher effort and instruction effectiveness using teacher accountability legislation data. *Journal of Human Capital, 7*(3), 230-273.

Assessing teachers: A conversation with Charlotte Danielson. (2012). *Principal, 91*(4), 26-27.

Bill and Melinda Gates Foundation. (2013a). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. Seattle, WA: Author.

Daley, G., & Kim, L. (2010). *A teacher-evaluation system that works*. Santa Monica, CA: National Institute for Excellence in Teaching.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford, CA: Stanford Center for Opportunity Policy in Education.

Darling-Hammond, L., & Ball, D. L. (1998). *Teaching for high standards: What policymakers need to know and be able to do*. National Commission on Teaching & America's Future.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job.* New York, NY: The Brookings Institute.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How 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. doi: 10.3102/002831213517130

Hill, H. C., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371-384.

Hinchey, P. H. (2010). *Getting teacher assessment right: What policymakers can learn from research*. Boulder, CO: National Education Policy Center. Retrieved February 15, 2019, from https://nepc.colorado.edu/publication/getting-teacher-assessment-right

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.

Koops, J. B., & Winsor, K. A. (2006). Creating a professional learning culture through faculty evaluation. *The Journal of Education, 186*(3), 61-70.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee value added assessment system. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education, 46*(4), 440-455.

Marshall, M. (2005). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, *86*, 727-730.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timerperley, H., & Earl, L. (2014). State of the art--teacher effectiveness and professional learning. *School Effectiveness & School Improvement*, *25*(2), 231-256. doi: 10.1080/09243453.2014.885451

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

Reinking, A. K. (2015). Increasing accountability measures for early childhood teachers using evaluation models: Observation, feedback, and self-assessment. *Current Issues in Education, 18*(1), 1-10.

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

Roberson, S., & Roberson, R. (2009). The role and practice of the principal in developing novice first-year teachers. *Clearing House, 82*(3), 113-118.

Sartain, L., Stoelinga, S. R., & Brown, E. (2011). Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation. Chicago, IL: Consortium on Chicago School Research.

Shakman, K., Riordan, J., Sanchez, M. T., Cook, K. D., Fournier, R., & Brett, J. (2012). *An examination of performance-based teacher evaluation systems in five states. Issues and answers* (REL 2012-No. 129). Waltham, MA: Institute of Education Sciences, Regional Educational Laboratory Northeast & Islands.

Stumbo, C., & McWalters, P. (2011). Measuring effectiveness: What will it take? *Educational Leadership, 68*(4), 10-15.

Toch, T. (2008). Fixing teacher evaluation. *Expecting Excellence, 66*(2), 32-37.

Wayne, A. J., & Young, P. (2003). Teacher characteristics and student achievement grains: A review. *Review of Educational Research, 73*(1), 89-122.