Educ Asse Eval Acc (2014) 26:5–28
DOI 10.1007/s11092-013-9179-5

Teacher evaluation and school improvement: An analysis of the evidence

Philip Hallinger · Ronald H. Heck · Joseph Murphy

Received: 16 September 2013 / Accepted: 30 December 2013 /
Published online: 16 January 2014

Abstract In recent years, substantial investments have been made in reengineering systems of teacher evaluation. The new generation models of teacher evaluation typically adopt a standards-based view of teaching quality and include a value-added measure of growth in student learning. With more than a decade of experience and research, it is timely to assess empirical evidence bearing on the efficacy of this school improvement strategy. This paper examines the new generation of teacher evaluation along three lines of analysis: evidence on the magnitude, consistency, and stability of teacher effects on student learning, evidence on the impact of teacher evaluation on growth in student learning, and literature from the sociology of organizations on how schools function. Although the trend towards focusing on teacher evaluation is increasingly evident internationally, most of the empirical research evaluated in this paper is from the USA. This critical evaluation of the empirical literature yields two key conclusions. First, we conclude that the policy logic supporting this reform remains considerably stronger than the empirical evidence. Second, we suggest that alternative improvement strategies may yield more positive results and at a lower cost in terms of staff time and district funds.

Keywords Teacher evaluation · Educational effectiveness · Teacher effectiveness

Efforts to improve teacher quality through performance evaluation have assumed an increasingly high profile position in the platform of education reforms undertaken by governments internationally (Atkinson et al. 2009; Flores 2012; Gray et al. 1995;

P. Hallinger (✉)
Hong Kong Institute of Education, 10 Lo Ping Rd., Tai Po, N.T., Hong Kong, China
e-mail: hallinger@gmail.com

R. H. Heck
College of Education, University of Hawaii-Manoa, 1776 University Avenue, Honolulu, HI 96822, USA
e-mail: rheck@hawaii.edu

J. Murphy
Vanderbilt University, Box 414, 230 Appleton Place, Nashville, TN 37203-5721, USA
e-mail: joseph.f.murphy@vanderbilt.edu

Harvey 2005; Hopkins and Stern 1996; Leithwood and Earl 2000; Liu and Zhao 2013; Odden and Wallace 2008; Reynolds et al. 2003; Robinson and Timperly 2007; Sanders and Rivers 1996; Skedsmo 2011; Thomas 2001). This raises two questions: (1) Why has this happened? (2) Is intensifying the focus on teacher evaluation likely to result in substantial improvements in the learning outcomes of students?

The answer to the "why" question can be traced to an emerging consensus on the hallmark place of teaching quality in school success (e.g., Hanushek 2010; Hattie 2009; Lewis 2008; Louis et al. 2010; Odden and Wallace 2008; Sanders and Horn 1994). A growing body of international research confirms a direct relationship between teacher quality/effectiveness and student learning (e.g., Goldhaber and Anthony 2007; Creemers and Kyriakides 2008; Hanushek 2010; Hattie 2009; Kyriakides et al. 2009; Hattie 2009; Liu and Zhao 2013; Sanders et al. 2005; Wright et al. 1997).

This broadening consensus on the importance of teaching quality has emerged during an era of increasing educational accountability throughout the world (Atkinson et al. 2009; De Fraine et al. 2002; Leithwood and Earl 2000; Liu and Zhao 2013; Flores 2012; Walker and Ko 2011). Over the past several decades, education policy has gradually shifted from holding schools accountable for policy compliance to accountability for learning outcomes (Atkinson et al. 2009; Hamilton et al. 2008; Harvey 2005; Kelly and Downey 2010). Within this changing global context, the search for more powerful strategies aimed at improving student performance has led policymakers and system leaders to experiment with new models of teacher performance evaluation (Attinello et al. 2006; Danielson 2007; Kimball et al. 2004; Milanowski et al. 2004; Sanders et al. 2005; Wilson et al. 2014). Indeed, after a period of relative neglect and pessimism (see Barth 1986; Bridges 1990; Medley and Coker 1987; Wise et al. 1985), policymakers increasingly view teacher evaluation as a potentially powerful means of filtering out poor-quality teachers and stimulating instructional improvement among teachers at large (Gates Foundation 2013; Gray et al. 1995; Heneman and Milanowski 2007; Odden and Wallace 2008).

The second question represents the focus of this paper. More specifically, we ask: "Is the reallocation of school resources (e.g., teacher and administrator time, development and maintenance of documentation systems, financial rewards) towards teacher evaluation likely to provide a robust pathway for school improvement?" Here, in spite of policymakers' fervent embrace, the evidence is less clear. Indeed, a perusal of global commentary suggests that the movement to intensify the focus on teacher evaluation needs more scrutiny than it has received to date (e.g., Darling-Hammond et al. 2012; Kelly and Downey 2010; Murphy et al. 2013). Although the efficacy of performance management is grounded in an appealing "managerial logic" (Ball 2003; Harvey 2005), we argue that this school improvement intervention should be assessed in light of empirical evidence on its effectiveness.

In this paper, we examine empirical evidence in an effort to understand the extent to which the "new generation" of teacher evaluation schemes is likely to improve the quality of teaching and learning in schools. More specifically, we critically evaluate three types of evidence.

1. Empirical evidence on the magnitude, consistency, and stability of teacher effects on student learning

2. Empirical evidence on the impact of teacher evaluation on growth in student learning
3. Literature from the sociology of organizations on the organizational processes that bears upon use of teacher evaluation as a vehicle for school improvement

Large amounts of money have been invested in the development and implementation of new systems of teacher evaluation over the past 15 years (Danielson 2007; Gates Foundation 2013; Heneman and Milanowski 2007; Kelly and Downey 2010; Millman 1997). Thus, we believe that this updated review of empirical research offers a timely assessment of this intervention based on observed results. The findings should be valuable in terms of shaping future directions in both policy and practice.

1 Conceptual perspective

Although we acknowledge that "teacher performance evaluation" and "instructional supervision" share some common features, we join other scholars (Castetter 1976; Duke 1990; Millman 1981, 1997; Popham 1988) by treating them as conceptually distinct constructs. We define teacher evaluation as "the formal assessment of a teacher by an administrator, conducted with the intention of drawing conclusions about his/her instructional performance for the purpose of making employment decisions" (Castetter 1976). This definition highlights the "personnel function" of teacher evaluation (Bridges 1990).

In contrast, we refer to instructional supervision as growth-oriented coaching conducted by administrators, supervisors, or peers. Instructional supervision employs a process of observation and feedback aimed solely at developing teaching capacity. Data gathered during the supervision process are not employed for employment decisions (see Attinello et al. 2006; Duke 1990; Ellett and Teddlie 2003; Leithwood 2001; Millman 1981, 1997; Popham 1988; Robinson and Timperly 2007; Showers 1985).

The logic of using teacher evaluation as a strategy for school improvement is predicated on the strength of the causal relationship between teacher quality and growth in student learning (Gates Foundation 2013; Milanowski et al. 2005; Odden and Wallace 2008). More specifically, researchers make two key relevant assertions.

- Variations in the quality of teachers are associated with differences in the learning gains of students (e.g., Sanders and Horn 1994).
- Teaching quality is subject to reliable and valid measurement capable of distinguishing the performance of teachers with respect to the achievement of their students (e.g., Danielson 2007; Hanushek 2010; Milanowski 2004a; Rockoff and Speroni 2010; Wright et al. 1997).

Drawing upon these assertions, policy advocates have proposed that teacher evaluation *can* and *should* be employed as a tool for managing teacher quality (Gates Foundation 2013; Odden and Wallace 2008; Toch and Rothman 2008). The logic underlying teacher performance as a school improvement strategy can be represented as type of "causal chain" (see Fig. 1). Advocates propose that teacher evaluation will

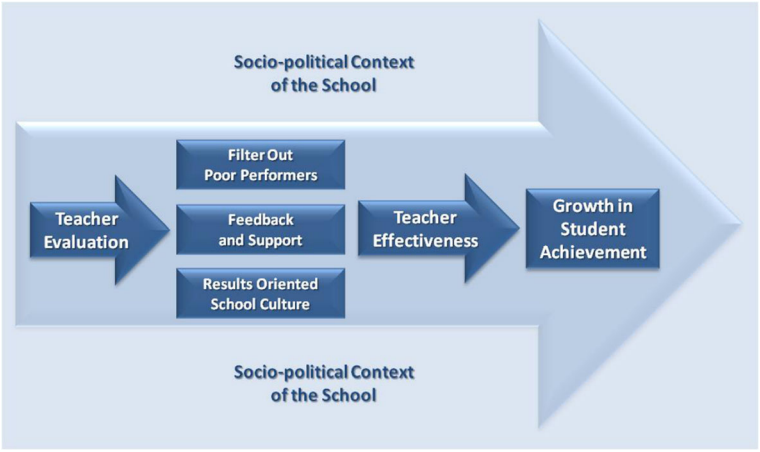

Fig. 1 Theory of action underlying teacher evaluation and school improvement

positively impact growth in student learning outcomes through three interrelated paths (Danielson 2007; Heneman and Milanowski 2007; Koppich and Showalter 2005).

1. First, performance evaluations should be capable of "weeding out" the weakest teachers, those failing to produce consistently positive effects on student learning (Bridges 1990; Gleeson and Husbands 2003; Harvey 2005; Heneman and Milanowski 2007; Koppich and Showalter 2005; Odden and Wallace 2008).
2. Second, performance evaluations will provide teachers with meaningful feedback, thereby resulting in improved quality of instruction and growth in student learning (e.g., Gates Foundation 2013; Heneman and Milanowski 2007; Odden 2004; Wright et al. 1997).
3. Third, teacher evaluation will contribute to development of a results-oriented school culture that will support a broader set of policy interventions designed to foster quality in teaching and learning (De Fraine et al. 2002; Ellett and Teddlie 2003; Hopkins and Stern 1996; Odden 2004; Reynolds et al. 2003).

These propositions are reflected in two key features that distinguish the new generation of teacher evaluation models that have emerged over the past 15 years. First, these evaluations of teachers are grounded in "observable standards" designed to enhance the quality of judgments made concerning teacher effectiveness (Danielson 2007). Administrators collect data on teacher classroom behavior through "low inference" classroom observations and compare the results against stated standards (Danielson 2007).

Second, evaluations systematically incorporate data on the achievement of the teacher's students over the preceding year (e.g., Danielson 2007; Gates Foundation 2013; Milanowski 2004a; Kelly and Downey 2010). Especially popular among this new generation of teacher evaluation models is the use of "value-added measures" (VAMs) of student gains in learning for that year. VAMs are proposed to represent the individual teacher's impact on the learning of his/her students during the prior year and

describe one important dimension of the teacher's "effectiveness" (Danielson 2007; Gates Foundation 2013; Kelly and Downey 2010; McCaffrey et al. 2003; Milanowski 2004a).

This represents a controversial departure from prior approaches to teacher evaluation that were typically "procedural" in nature. Past teacher evaluation models typically employed "high inference" methods of performance assessment (e.g., checklists) and seldom, if ever, included data on student achievement (Bridges 1990; Duke 1990; Millman 1981, 1997; Wise et al. 1985). We characterize this shift in focus as "controversial" because of the long-standing finding that the greatest proportion of variability in student learning outcomes is determined by family background (Coleman et al. 1966). That is, it was previously deemed unfair to hold teachers accountable for a process–product relationship between teaching and learning that was determined *to such a large degree* by factors outside the control of the individual teacher.

The policy logic of teacher evaluation as a "school improvement strategy" is based on several interrelated assumptions. First, it is assumed that the "magnitude of the teacher effect" on growth in student learning is sufficient to warrant inclusion of student achievement data in performance evaluations. Second, this logic assumes that the measurement tools used in these evaluations are capable of reliably capturing and differentiating the impact of teachers on growth in student learning (Bridges 1990; Kelly and Downey 2010; Latham and Wexley 1981). Third, it is assumed that the application of this approach to teacher evaluation will produce consistent and sustainable improvements in the quality of teaching and learning (Darling-Hammond et al. 2012). Finally, it is assumed that this intervention will justify the considerable investment of human and financial resources required for its effective implementation (e.g., Bridges 1990; Range et al. 2011; Skedsmo 2011). Our multi-pronged analysis of relevant literature will evaluate the validity of this policy logic.

2 Method

This paper employs a research review methodology (Gough 2007; Hallinger 2013) focusing on three distinct literatures. These consisted of empirical evidence on teacher effects, empirical evidence on the implementation and effects of new generation teacher evaluation models, and evidence on school organization and school improvement. The review was guided by a framework for conducting systematic reviews of research (Hallinger 2013). However, the extent to which the components of systematic review were employed varied across the three literatures. Thus, for example, while the first two reviews sought to identify and review relevant empirical literature, the third review had a broader brief. It sought to place the results of the first two sections in historical and organizational perspective.

As indicated above, the paper is actually comprised of reviews of three related but distinct literatures. The first was comprised of a subset of the quantitative empirical literature on teacher effectiveness. This subset consisted of studies that examined teacher effects on growth in student learning outcomes using "value-added" data. We chose this subset of the broader teacher effectiveness literature because it represents one of the two foundation blocks for the new generation of teacher evaluation models. The second literature consisted of empirical studies of the implementation and/or impact of

new generation teacher evaluation models. Finally, we examined literature on school organization and school improvement.

In the first domains, the authors used Google Scholar™ to search for relevant studies. Our search approach was "exhaustive" rather than "bounded" (Hallinger 2013), since we knew that the number of empirical studies in these domains would be limited. Consistent with an exhaustive search, we included all relevant sources including journal articles, book chapters, doctoral dissertations, conference papers, and working papers in the database of studies. The search for relevant literature on school organization and improvement relied heavily on prior reviews of this literature conducted by the authors themselves.

For each relevant study, we extracted information concerning conceptual and methodological characteristics as well as substantive results. Thus, we not only paid attention to the pattern of findings (e.g., effect sizes), but also to research designs and conceptual models. This information was extracted, stored, and organized for subsequent analysis.

The mode of analysis employed in the study consisted of critical evaluation of the literature rather than quantitative analysis or meta-analysis. Critical evaluation entailed examining the pattern of conceptualization, methodology, and findings across the sets of studies in order to discern patterns. As we elaborate later in the paper, the nature of these literatures made it essential to employ a critical lens rather than integrative algorithms. The interplay of conceptual models, statistical models, and findings must be considered in concert.

3 Assessing the evidence

We begin our analysis by focusing on evidence concerning teacher effectiveness. Then, we proceed to examine empirical evidence from studies that have assessed the impact of the new generation teacher evaluation models on student learning. Finally, we place this policy intervention in the organizational context of schools. This analysis will not only evaluate the empirical evidence on teacher evaluation as a school improvement strategy, but also offer perspective on the organizational conditions required in order for teacher evaluation to fulfill the causal chain implied in Fig. 1.

3.1 Empirical evidence of teacher effects on learning outcomes

Researchers have asserted that a significant proportion of variance in student learning can be traced to differences in the quality of teachers (Goldhaber 2002; Hanushek 1992, 2010; Hanushek and Rivkin 2010; Milanowski et al. 2005; Rockoff 2004; Rivkin et al. 2000, 2005; Rowan et al. 2002; Sanders and Horn 1994; Sanders and Rivers 1996; Toch and Rothman 2008; Wright et al. 1997). The differential effectiveness of teachers, however, has been described and studied in different ways (Ellett and Teddlie 2003). Although the differences among these approaches may appear "technical," we assert that they can have a significant impact on the size of the effect on growth in student learning that is attributed to an individual teacher.

The models associated with using VAMs for studying teacher effects are a subset of analytical models that examine patterns of student growth in learning over time. They

differ from traditional approaches of assessing teacher effectiveness by inferring the teacher's effectiveness from a mathematically adjusted estimate of students' gains in achievement over the course of a school year. The initial "two-level" (i.e., teacher level and student level) approach was developed by Sanders et al. (1994, 1996) during the mid-1990s.

The VAM approach uses the prior test scores of students to "condition the learning effect" for the current year. After controlling for prior learning, the remaining "gain score" is proposed as an indicator of the current teacher's "effect" on the student's learning (Rivkin et al. 2000, 2005; Sanders and Horn 1994; Sanders and Rivers 1996; Wright et al. 1997). Researchers place teachers into categories of effectiveness (e.g., quartiles) based on their students' gains in achievement and have demonstrated that different categories of teachers were associated with different levels of "teacher effectiveness" in producing student learning growth over time (Wright et al. 1997). Based on these findings, Wright et al. concluded that "teacher evaluation processes should include, as a major component, a reliable and valid measure of a teacher's effect on student academic growth over time" (Wright et al. 1997, p. 66).

This approach to studying teacher effectiveness has succeeded in identifying differences in student learning across various categories of teachers. We caution, however, that these empirical results represent only a modest empirical estimate of teachers' contribution to student learning within research designs that do not measure student growth in an optimal manner. The bulk of supporting evidence for the VAM approach to studying teacher effectiveness has been based on "two-level research designs" that only consider students as nested within classrooms. This type of design can mask the potential effects of schools in defining teachers' work in classrooms and may incorrectly attribute the effects of missing school-level variables (e.g., school academic and social organization) to teachers.

We illustrate this problem by reference to several multilevel studies that accounted for three levels (i.e., schools, teachers, students) rather than two levels of analysis (i.e., teachers and students). This type of VAM represents an expanded version of the two-level approach (e.g., Sanders and Horn 1994; Sanders and Rivers 1996). It explicitly seeks to account for variance in student achievement that is due to the organizational structure of schools—for example, the nonrandom grouping of students within classrooms and classrooms within schools—and to account for classroom and school covariates in the estimation of teacher effects on student learning.

In early studies, Scheerens and Bosker (1997) estimated that roughly 15–20 % of variance in student achievement outcomes was due to features of schools, 10–15 % to teachers, and 60–70 % to students at any given point in time. Goldhaber (2002) reported similar estimates of variance in student achievement across student, classroom, and school levels. He found that about 79 % of variation was accounted for by student characteristics, about 8.5 % of the variance was due to differences among teachers, and roughly 12.5 % was accounted for by differences in the conditions presented by school organization and capacity.

Rowan et al. (2002) also noted considerable differences in variance attributed to classrooms (and teachers); that is, roughly 12–28 % for reading and math in their initial models without covariates. They further noted that after adjusting for covariates (i.e., prior-year scores, classroom demographics, school demographic

composition), initial classroom variance was reduced to between 4 and 16 %, depending on whether the outcome was reading or math. They also observed that student background variables had different effects on estimated student annual gains. In their "cross-classified" models, Rowan et al. (2002) attributed much larger gains to individual teachers than in their simple gain score models.

The cross-classification of students within different teachers allows the estimation of current achievement by including combinations of the student's past and current teachers at the classroom level of the analysis. The effects of previous teachers do not have to be assumed to be constant. One limitation, however, is that when estimating student gains during a particular grade level, only the current teacher actually contributes to gains and not the previous teacher or teachers (McCaffrey et al. 2003). Moreover, student gains (as well as growth trajectories) can be challenging to estimate accurately, especially when a student has multiple teachers (Darling-Hammond et al. 2012). Hence, the effects attributed to a specific teacher can vary considerably depending on the grade level at which growth is examined.

In Table 1, we provide a simple illustration of this problem of attributing greater variance in outcomes to teachers in two-level models versus including teachers nested within their schools for four commonly used VAMs for estimating teacher effects on elementary students' reading scores. In this example, we used a sample data set comprised of approximately 10,000 students cross-classified over 2 years within 1,000 teachers and nested within 160 elementary schools.

- Model 1 is cross-sectional and considers proportions of variability in fifth grade reading scores due to students, teachers, and schools (similar to Scheerens and Bosker 1997 and Goldhaber 2002).
- Model 2 incorrectly assigns the observed school variance in reading to the teacher level of the model.
- Model 3 is cross-classified and adds possible variability due to the previous year's teachers, but incorrectly ignores the school-level variability.
- Model 4, which is similar to the Rowan et al. (2002) cross-classified models, considers the variance in reading scores due to students, teachers cross-classified at level 2 (but does not make the assumption that the first teacher effects must persist at year 2), and schools at level 3.

Table 1 Estimates of variance proportions for three- and two-level reading models

Parameter	Model 1	Model 2	Model 3	Model 4
Intercept [school]	0.088*			0.079*
Intercept [teach year 2]	0.068*	0.154*	0.132*	0.070*
Intercept [teach year 1]			0.049*	0.023
Residual [student]	0.854*	0.846*	0.819*	0.828*
Total variance	1.000	1.000	1.000	1.000

*$p<0.05$

The results illustrate how omitting hierarchical levels (such as the school or classroom) and estimating different types of nested multilevel models change the allocation of variance in ways that may affect the estimation of teacher effects. As noted earlier, the studies in this research domain do not typically nest teachers within schools or include relevant school-level factors in their models (e.g., Hanushek 2010; Hanushek and Rivkin 2010; Rivkin et al. 2005; Rockoff 2004; Sanders and Horn 1994; Sanders and Rivers 1996; Wright et al. 1997). As McCaffrey et al. (2003) concluded, the omission of school-level variables results in biased estimates of teachers' contributions to student learning.

Clearly, teachers do not work in isolated, individual environments within schools. This is especially so in secondary school settings, where students' learning unfolds within complex sorting process that is shaped by courses, peers, teachers, and schedules during the instructional day (Garet and Delany 1988; Hallinger et al. 2014). Examining this complexity in detail reveals several distinct socio-curricular paths through which students experience high school. With respect to the evaluation of secondary school teachers, if we ultimately determine that a particular student's 50 or more high school teachers contribute 0.2 of a standard deviation (SD) in terms of growth on a standardized test against the "average" student, is it either feasible or justifiable to evaluate and compensate them all differentially by their weighted contribution to student learning?

Both theoretical and technical advantages, therefore, result from including schools as an analytic level within models of teacher effectiveness. Aside from gaining an understanding about how important school processes (e.g., strategic instructional improvement, academic press, school leadership) may condition teachers' work, student composition variables that cluster at the school level (e.g., SES, language background) will not be confounded with teacher effects. Instead, the effects of the clustered student characteristics will be correctly identified at the school level (McCaffrey et al. 2003).

Including the school level also allows the consideration of differences in average teaching effectiveness (as well as variability in effectiveness within each school) to be included in the models. As Rothstein (2009) has noted, differences in teacher effectiveness that may be present at the school level can also affect the process of student assignment to teachers, which can bias estimates of teacher effects. Rowan et al. (2002) also make the point that at the elementary level, personnel seem to allocate pupils to more and less effective teachers based more on chance rather than a systematic process.

Along with other reviewers of this body of research (e.g., Ellett and Teddlie 2003; Hattie 2009; Lewis 2008; Louis et al. 2010; Walberg 2011), we conclude that teachers do have a measurable effect on student learning. Reported standardized teacher effects within schools tend to be in the range of 0.1 to 0.2 SD in test scores (e.g., Aaronson et al. 2007; Borman and Kimball 2005; Rivkin et al. 2000, 2005; Rothstein 2009; Rowan et al. 2002). The presence of these effects can depend on the particular test and other features of the statistical models employed in the research. Importantly, these effect sizes across studies in many different contexts are not sizeable enough to conclude that teachers bear complete responsibility for student learning in a manner that would suggest employing or not employing them on the basis of a value-added score. Thus, we caution that both the magnitude and sustainability of effects of teacher-related variables on student learning outcomes remain inconsistent and at times overstated (Bressoux and Bianco 2004; Baker et al. 2010; Darling-Hammond et al.

2012; McCaffrey et al. 2003; Rothstein 2009). Nonetheless, even if we acknowledge the existence of teacher effects on student learning, we still have not resolved the question whether assessments of teacher performance can be used to leverage instructional improvement and, if so, how to best utilize them. In the next section of the article, we examine empirical evidence on the effects of new generation teacher evaluation models that make use of value-added data on student learning.

3.2 Empirical evidence on the effectiveness of teacher evaluation

Principals have been evaluating teachers for over a century. Not surprisingly, a good deal of descriptive data and prescriptive opinion on the efficacy of teacher evaluation has accumulated over time (see Bridges 1967, 1990; Danielson 2007; Grotke 1953; Medley and Coker 1987; Millman 1981, 1997; Showers 1985; Stiggins and Duke 1988; Wise et al. 1985). Indeed, during the twentieth century, scholars proposed various approaches to teacher supervision and evaluation. However, empirical efforts to explore the relationship between teacher evaluation by principals and student learning outcomes were few and far between (Hallinger and Heck 1998; Wise et al. 1985). This can be explained, in part, by the fact that earlier generations of teacher evaluation models were not *explicitly* linked to the assumption that this administrative practice would produce a measureable impact on student learning (Millman 1997).

As suggested, however, by our discussion of teacher effectiveness, the past decade has yielded a new generation of teacher evaluation models (Danielson 2007). Fortunately, implementation of this new generation of teacher evaluation has been accompanied by a demonstrable increase in the number of empirical studies of impact. In this section, we assess the trend of findings from the major studies that comprise this body of research (Table 2).

Webster and Mendro (1997) studied one of the first efforts to implement value-added teacher evaluation as part of the Dallas (TX) School District's efforts to identify effective schools. The initiative provided monetary rewards for teachers within top-performing schools in several categories of elementary, middle, and high schools (as opposed to providing rewards based on individual teacher performance). Subsequently, teacher effectiveness indices based on two-level models were devised using a value-

Table 2 Major studies that comprise this research

Author/year	Location	Model	Grades	Years	Findings	Evaluation
Webster and Mendro (1997)	Dallas, TX	2-level				
Mendro (1998)	Dallas, TX	2-level				
Bembry and Schumacker (2002)	Dallas, TX	2-level				
Borman and Kimball (2005)	Reno, NV	2-level				
Kimball et al. (2004)	Washoe City, NV	2-level				
Milanowski (2004a, b)	Cincinnati, OH	2-level				
White (2004)	Coventry, RI	2-level				
Kimball and Milanowski (2009)	1 district western US state	2-level				
Gates Foundation (2013)	7 school districts	2-level				

added approach. Webster and Mendro (1997) noted that principals had some difficulty in using the indices for evaluation purposes and recommended caution in applying the results.

They also noted a number of technical problems associated with implementing more sophisticated three-level analyses sought to account for school covariates. A follow-up study by Mendro (1998) noted identifiable effects persisting into the future for students taught by a highly effective versus a noneffective teacher. However, interpretation of the findings of this study was subject to alternative conclusions, with Bembry and Schumacker (2002) cautioning against using value-added measures to evaluate individual teachers.

Borman and Kimball (2005) studied a sample of 400 teachers and 7,000 students in one school district in Reno, NV. Their goal was to assess whether the standards-based evaluation system helped close the achievement gap among students of different socioeconomic backgrounds. Using a two-level model (i.e., students nested within teachers), their results showed higher mean achievement in classrooms taught by teachers of higher quality; however, the actual magnitude of differences was quite small. They concluded:

> This analysis suggests that teacher quality, as defined and applied in the evaluation system of one school district, may not show reliable relations to closing achievement gaps between poor and more advantaged, minority and nonminority, and low- and high achieving students. The implications for the evaluation system are important, especially if a key component of teacher quality is an ability to close achievement gaps. (Borman and Kimball 2005, p. 18)

Kimball et al. (2004) conducted a wider scale study of a standards-based teacher evaluation system in Washoe County, Nevada in which they sought to understand if, "teachers who score well on such evaluation systems also help produce higher levels of student learning" (p. 56). This research employed a two-level model to examine the relationship between teacher evaluation results and student gains in achievement in reading and math. The results were mixed, and the overall findings offer little in the way of empirical evidence supporting the efficacy of teacher evaluation.

Milanowski (2004b) examined the implementation of a standards-based evaluation system in Cincinnati. Consistent with the other studies described above, he used a two-level model to determine the nature of the relationship between the evaluation scores of teachers and VAMs of student learning in grades 3 through 8. The author noted that the school system's administrators, "want to be justified in inferring that teachers with high scores are better *performers*, defined as producing more student learning" (Milanowski 2004b, p. 39). Although the study yielded some positive results, there was an inconsistent pattern of teacher evaluation results and student gains across grades and subjects. In spite of these mixed results, the researcher concluded that the "moderate level of criterion-related validity" (p. 49) was sufficient to support the use of student achievement data in the evaluation of teachers.

White (2004) conducted a study in Coventry, Rhode Island that sought to "describe the relationship between a teacher's overall evaluation score and his or her students' achievement, while controlling for prior achievement, in order to determine the criterion-related validity of the evaluation scores" (White 2004, p. 3). He followed a

similar two-level approach as Milanowski's (2004b) study in analyzing value-added achievement data in reading and math from 3,617 students and teacher evaluation data for 173 teachers in four elementary school grades and two school years. White's results "indicate[d] a small overall correlation in reading (0.240) and essentially no correlation in math (0.032). The results also indicate rather large fluctuations in correlations between years and across subjects and grade levels" (p. 6). Again, the overall pattern of results provide weak empirical evidence supporting the efficacy of teacher evaluation in elementary schools.

Kimball and Milanowski (2009) conducted a study in which they examined variation in the ratings of teachers within another school district that had implemented standards-based evaluation. Their study focused on the validity of the ratings obtained by different principals in the study and then sought to relate these ratings to value-added achievement results of students. The findings reflected an uncertain relationship between the evaluations and value-added measures of student learning. The lower than expected validity of the ratings (i.e., relationship between the ratings and VAMs) was noted by the authors as a cause of concern. More specifically, they suggest that the low validity was attributable to a set of "complex and idiosyncratic" factors that appeared to bear upon principals' decision-making. The authors concluded:

> We had hoped that we could identify evaluator practices associated with higher validity, which districts could then use to train evaluators to follow. Although disappointing, our failure to find such practices is important because it shows the complexity in identifying and assuring the use of good evaluation practice . . . If policy makers and program designers want evaluation scores to be more highly related to some criterion such as student achievement, it will take more than specific rubrics and basic training of evaluators in the process to achieve a strong relationship. (Kimball and Milanowski 2009, p. 65)

The most recent large-scale effort to assess the efficacy of this approach to teacher evaluation is represented in the Gates Foundation-funded *Measures of Effective Teaching* study (Gates Foundation 2013). This study employed a two-level research design similar to those discussed above. The study differed largely in terms of the size (seven large school systems and 3,000 teachers), grade levels, and length (3 years of data) of the study. Recent reports have suggested that the findings of this highly publicized study affirm a positive relationship that is stable from year to year between the assessed quality of teachers and student gains in learning (Gates Foundation 2013).

Nonetheless, this methodology of this study not only suffers from the same methodological limitations discussed in the prior section, but also from an optimistic and overstated interpretation of the actual findings. For example, the methodology used by the researchers did not make it possible to assess the stability of results for individual teachers. As Glass (2013) has pointed out, the researchers analyzed the stability of results for groups of teachers rather than individual teachers.

> Just because the average of VAM scores for 150 teachers will agree with next year's VAM score average for the same 150 teachers gives us no confidence that an individual teacher's VAM score is reliable across years. In fact, such scores are not—a fact shown repeatedly in several studies. So we aren't going to fire groups

> of 150 teachers arbitrarily lumped together who might have low VAM scores, nor pay big bonuses to the high VAM group. Nor are we going to fire those teachers whose Language Arts VAM score is low, because the odds are substantial that the same teachers' Math VAM score might be average or even above. (Glass 2013)

Glass' comments on the MET study again highlight the limitations of the evidence base on using VAMs in practice. The use of VAMs in evaluating the performance of individual teachers has come under criticism for a variety of reasons, some of which have already been discussed. Darling-Hammond et al. (2012) and Darling-Hammond and Youngs (2006) summarized three key limitations of using value-added measures for the purposes of teacher evaluation:

1. Value-added models of teacher effectiveness yield inconsistent patterns of results for individual teachers over time, thereby calling into question their validity for the purposes of performance appraisal.
2. Teachers' value-added performance is affected by the students assigned to them in a given year, thereby calling into question the transparency and fairness of using value-added measures of student learning in evaluations.
3. Value-added ratings are unable to disentangle the many other influences that contribute to student progress, thereby providing an incomplete and distorted measure of an individual teacher's effectiveness (Darling-Hammond et al. 2012, pp. 9–11).

When data on staff performance are intended for use in making personnel decisions, it is incumbent upon system designers and administrators to demonstrate that the relevant instruments and methods yield results that meet accepted standards of validity (Latham and Wexley 1981). Messick (1994) further observed: "The consequential basis of test validity includes evidence and rationales for evaluating the intended and unintended consequences of test interpretation and use in both the short and the long term. Particularly prominent is the evaluation of any adverse consequences for individuals and groups that are associated with bias in test scoring and interpretation or with unfairness in test use" (p. 21).

The empirical findings reported in this section of the paper present a pattern of weak, inconsistent, and unstable results with respect to the relationship between standards-based teacher evaluations and student learning gains across subject areas, grade levels, and intervals of time. We further note that all but one of the studies of the effectiveness of standards-based/VAM teacher evaluation (i.e., Gates Foundation 2013) were conducted at the elementary school level. This is not surprising since structural complexity makes it difficult to apply and validate VAM teacher evaluation models for use in secondary schools.

With these limitations in mind, we assert that that the current use of VAMs offers limited utility for the purpose of determining the effectiveness of an individual teacher working with a specific set of students. This limitation highlights the challenge of moving from research findings into the construction of policy solutions. Therefore, we conclude that standards-based teacher evaluation systems have to meet a standard of validity necessary for making personnel decisions. More broadly, we have yet to see compelling evidence that the implementation of

these system is yielding higher teaching quality and improved learning outcomes for students.

3.3 Indirect evidence on teacher evaluation as a school improvement strategy

The first two parts of this review offer, at best, weak support for investing in teacher evaluation as a strategy for school improvement. However, even if we accept that problems cited in the prior sections can be overcome, what is the likelihood that teacher evaluation can fulfill its aims of enhancing teacher quality and fostering more consistent growth in student learning? When we examine teacher evaluation as a strategy for school improvement, it must be assessed in relation to the effects of other alternative interventions, the financial costs of implementing designs that produce consistent results, and the possible negative consequences that attend its implementation (Hawley and Rosenholtz 1984; Murphy et al. 2013).

In this section, we examine "indirect evidence" on the efficacy of teacher evaluation drawn from the literatures on educational effectiveness (e.g., Creemers and Kyriakides 2008; Hanushek 2010; Kyriakides et al. 2009) and school improvement (e.g., Purkey and Smith 1983; Reynolds et al. 2000). If teacher evaluation "works," we might also expect to see it emerge in these broader-related literatures. However, we find that teacher evaluation has been and continues to be conspicuous by its absence in the following bodies of work: educational effectiveness (Creemers and Kyriakides 2008; Hanushek 2010; Kyriakides et al. 2009), school improvement (Reynolds et al. 2000; Teddlie and Reynolds 2000), instructional leadership (Hallinger and Heck 1998; Robinson et al. 2008), school restructuring (Murphy 1991), teaching change and development (Hattie 2009; Joyce and Showers 2002; Showers 1985), comprehensive school reform (Herman and Stringfield 1997), data-based decision-making (Supovitz and Klein 2003), school reform and change (Borman 2005; Fullan 2001), programs targeting at risk students (Reyes et al. 1999; Slavin et al. 1989), and turnaround schools (Leithwood et al. 2010; Murphy 2008). This does not mean that teacher evaluation could not be a driver of school improvement. Nonetheless, it seems prudent to be cautious when so little supporting evidence can be located.

In order to gain deeper insight into why this might be the case, we reexamine the theory of action that is powering the current focus on teacher evaluation. In doing so, we focus on the "organizational dynamics" of schools or what sociologists refer to as "occupational norms and workplace conditions" (e.g., Hamilton et al. 2008; Lortie 1975; Rosenholtz 1991). Over the past 50 years, numerous scholars have noted ways in which the organizational dynamics of schools can erode the theory of action on which the teacher evaluation equation is based (Barth 1986; Bidwell 1965; Blasé and Kirby 2009; Bridges 1967; Cuban 1988; Hamilton et al. 2008; Weick 1976).

For example, scholars and practitioners have found it difficult to reconcile the conflict between administrative efforts to intensify teacher performance evaluation and while engaging in development-oriented instructional supervision and development (Barth 1986; Blasé and Kirby 2009; Darling-Hammond et al. 2012; Joyce and Showers 2002; Marshall 1996; Popham 1988; Stiggins and Duke 1988). A deep and recurring theme in the instructional supervision and development literature emphasizes the potential costs of intensifying the focus on performance evaluation. Emphasizing the summative function of teacher evaluation may not only impede efforts to motivate

change in teacher behaviors, but also participation in complementary strategies aimed at building productive collaboration and community (Baker et al. 2010; Barth 1986; Hawley and Rosenholtz 1984; Joyce and Showers 2002; Popham 1988; Rosenholtz 1991; Stiggins and Duke 1988). In the words of Showers (1985):

> [N]othing could be farther from the atmosphere of coaching than is the practice of traditional evaluation. The norms of coaching and evaluation practice are antithetical and should be separated in our thinking as well as in practice. By definition, evaluation should not be undertaken concurrently with coaching. . . (p. 46)

Earlier reviews of teacher evaluation often highlighted the questionable validity of the tools that were placed in the hands of school principals (e.g., Bridges 1990; Medley and Coker 1987; Millman 1981; Wise et al. 1985). Cursory observations of classroom instruction offered a potentially biased and inadequate sample of teaching practice for the purposes of performance evaluation. Checklists focusing on the instructional and personal professional behavior also represented weak tools for differentiating the performance of teachers. This state of affairs took the starch out of teaching evaluations generally rendering them into procedural rituals that lacked meaning, legitimacy, and impact (Bridges 1990; Weick 1976; Wise et al. 1985).

In fairness, advocates propose that new generation teacher evaluation models employ a new and more robust set of tools (Danielson 2007; Gates Foundation 2013; Kimball and Milanowski 2009; Odden 2004; Toch and Rothman 2008; Rockoff and Speroni 2010). These "state-of-the-art tools" include a clear set of standards against which to benchmark teacher performance, more intensive observations of classrooms, validated instruments, and data on the learning gains of the particular teacher over the past year(s). Taken together, these tools are proposed to offer a more comprehensive and defensible means of assessing teacher performance for the purposes of instructional development as well as reward and sanction (Danielson 2007; Heneman and Milanowski 2007; Odden 2004; Odden and Wallace 2008; Rockoff and Speroni 2010).

Available evidence, however, suggests that equipping principals with the skills needed to operate the teacher evaluation machinery is more difficult than anticipated (Blasé and Kirby 2009; Darling-Hammond et al. 2012; Kimball and Milanowski 2009). For example, Kimball and Malinowski, arguably the most active empirical researchers studying implementation of the new generation of teacher evaluation approaches, admitted "disappointment" in the capacity of principals to fulfill the stated requirements in using these new tools. They concluded: "Our study does not dismiss will, skill, and context as potentially important factors in evaluation decision making, but it does illustrate the complexity in fully uncovering these factors" (Kimball and Milanowski 2009, p. 67). This conclusion reprises the disappointing experience of prior generations of reformers who sought to employ teacher evaluation as a lever for change in teaching and learning (Barth 1986; Bridges 1967, 1990; Camburn et al. 2003; Cuban 1988; Grotke 1953; Loup et al. 1996; Marshall 1996; Medley and Coker 1987; Millman 1981, 1997; Popham 1988; Showers 1985; Stiggins and Duke 1988; Wise et al. 1985).

Simply stated, principals have few incentives and many disincentives to invest their time in evaluating teachers (Bridges 1990; Cuban 1988; Marshall 1996). Moreover, regardless of the resources, rubrics, and requirements that policymakers and system

leaders thrust upon them, the appetite of principals for this task is unlikely to improve in the foreseeable future. There are fundamental reasons why principals tend to avoid the exercise of tight control over the pedagogical work of teachers (Barth 1986; Bridges 1990; Cuban 1988; Marshall 1996; Meyer and Rowan 1975; Weick 1976). Principals for the last century have understood that they can secure that support by providing teachers with some degree of autonomy over their classrooms (Cuban 1988). In turn, teachers have consistently traded off influence over school-level work for freedom in their classrooms (Barth 1986; Blasé and Kirby 2009; Cuban 1988; Duke et al. 1980; Marshall 1996).

As Bidwell (1965) observed, the organization of schooling provides teachers with professional discretion to make individual judgments regarding students' needs and abilities. This is necessary so that they can make needed adjustments in day-to-day instruction al activities. At the same time, however, the necessity for teachers to deliver a formal curriculum requires considerable uniformity and routinization in moving students sequentially from grade to grade and school to school within the education system. Balancing the needs for teacher autonomy and systemic uniformity, therefore, represents a primary task of school administrators. When viewed from this perspective, efforts to intensify teacher performance evaluation represent a "threat" to this normative balance. Thus, the potentially positive benefits of intensifying teacher evaluation must also be weighed against the potential negative costs of increased conflict between administrators and teachers.

Descriptions of how principals deal with this challenge in their daily work lives leave a powerful impression of complex countervailing pressures. Cuban (1988) referred to the persisting reluctance of principals to embrace these tasks as a type of genetic code embedded in the DNA of the principalship. Marshall (1996) described a state of ongoing frustration at his own inability, as a practicing principal, to penetrate the "force field of the classroom" despite his strongest intentions. The normative environment in which principals work is not going to change simply through appeals to rationality and exhortations from policymakers. Both organizational norms and structures must change before assertions concerning the power of teacher evaluation to leverage improvement can be taken seriously (Cuban 1988; Hamilton et al. 2008).

In addition, as noted, the new generation of teacher evaluation requires the time-intensive use of low inference methods of teacher observation and feedback. It is difficult to see how sufficient time and energy of school administrators can be infused into teacher evaluation to make it a viable tool in ratcheting up instructional quality. An unrealistically wide span of control already limits the total amount of time available for principals to engage in classroom supervision activities (Barth 1980; Bidwell 1965; Bridges 1990; Camburn et al. 2003; Marshall 1996). With this limitation in mind, researchers find that when principals do engage in instructional leadership, they tend to focus on school-wide rather than classroom-specific strategies (Hallinger and Heck 1998; Leithwood et al. 2010; Louis et al. 2010; May and Supovitz 2011; Murphy 2008; Robinson et al. 2008; Sebastian and Allensworth 2012; Spillane et al. 2009).

It is also the case that time for exercising instructional leadership must be balanced against a variety of competing managerial, organizational, and community leadership activities (Barth 1986; Hallinger and Murphy 2012; Murphy et al. 1987). Recent analyses of principal work time and tasks indicate that the average American principal spends an average of 18 % (a high estimate) of the work week engaged in managing

instruction and curriculum (see Horng et al. 2010; May and Supovitz 2011; Hallinger and Murphy 2012; Spillane et al. 2009). And only about 3 % of their work time is spent on teacher evaluation. These numbers are largely unchanged despite 30 years of concentrated efforts to increase them (Murphy 1990).

Data obtained from the International Education Assessment indicate widely varying amounts (and percentages) of time allocated to different management tasks by principals from country to country (Lee and Hallinger 2012). Nonetheless, these data suggest that the USA represents "an optimistic scenario" with respect to the time that principals allocate to instructional leadership (see Lee and Hallinger 2012). This role tends to consume an even smaller portion of the principal's work week in most other countries.

This review of the organizational literature was undertaken with two aims in mind. First, we wished to determine whether evidence could be found to support the use of teacher evaluation as a school improvement strategy from related literatures. We found no such evidence.

Second, we noted that teacher evaluation is a kind of school improvement intervention and, as such, does not take place in isolation. Therefore, we sought to understand how the implementation of new generation teacher evaluation models might "fit" into the task structure of school leadership and school improvement. Our assessment of these organizational dynamics makes it difficult to discern how formal teacher evaluation systems will lead to improved quality of instruction in schools. If they do not, then they are unlikely to have a positive impact on student learning in particular and school improvement in general.

4 Conclusion

Teacher evaluation has been reinvented and repositioned as a solution for improving teacher quality several times in the past (Cuban 1988; Gray et al. 1995; Grotke 1953; Millman 1981, 1997; Musella 1970; Popham 1988; Tyack 1974; Wise et al. 1985). Interest in teacher evaluation as a policy solution reemerged internationally during the late 1990s in concert with research that highlighted the impact of teacher quality on growth in student learning (e.g., Sanders and Horn 1994; Wright et al. 1997). In the context of increasing system demands for accountability, this led to experimentation with new designs for teacher evaluation in the USA (e.g., Danielson 2007; Ellett and Teddlie 2003), UK (e.g., Crosnoe 2011; Gray et al. 1995; Harvey 2005; Kelly and Downey 2010; Reynolds et al. 2003), Europe (Ball 2003; De Fraine et al. 2002; Flores 2012), and Asia (Liu and Zhao 2013; Walker and Ko 2011). This latest generation of teacher evaluation is often distinguished by a standards-based view of effective teaching combined with value-added measures of growth in student learning.

After more than a decade of implementing new generation models of teacher evaluation, this review examined evidence of results. Our review of three related literatures found that the "policy logic" driving teacher evaluation remains considerably stronger than empirical evidence of positive results. More specifically, the review found the following.

- Literature on the new generation of teacher evaluation is characterized by overly optimistic interpretations of the underlying literature and a tendency to overlook

important limitations of the research designs used in these studies of teacher effectiveness. In particular, most existing models fail to take into account other important school-level factors (e.g., nonrandom distribution of students) that can bear upon the efforts of individual teachers within and across schools.

- Efforts to translate academic research on teacher effectiveness into practical tools for monitoring the performance of individual teachers fail to meet the technical and administrative requirements needed for this professional task. This is especially evident at the secondary school level where students learn under the guidance of multiple teachers, making it even more challenging to tease out their differentiated effects on growth in learning.
- There is remarkably little evidence that associates the new generation of teacher evaluation with capacity development of teachers or more consistent growth in the learning outcomes of students. Indeed, research has highlighted the complexity of achieving the desired "fidelity of implementation" of the new teacher evaluation models in schools.
- A broader reading of related literatures on educational effectiveness and school improvement finds little support for the belief that teacher evaluation represents a high impact school improvement strategy.
- Finally, our analysis surfaced numerous reasons for why the administrators responsible for teacher evaluation find it difficult at best and counter-productive at worst to intensify their efforts at teacher evaluation.

We also noted that teacher evaluation is not implemented in a policy vacuum. The efficacy of teacher evaluation as a policy strategy should be assessed in relation to other alternatives. Many leadership-related initiatives can have significant effects on student learning, *even if they do not directly target the quality of teaching*. Examples include establishing strong academic mission with challenging organizational goals (Cotton 2000; Hallinger and Heck 1998; Hawley and Rosenholtz 1984; Robinson et al. 2008), enhancing student opportunity to learn (Balfanz and Byrnes 2006; Harris and Herrington 2006; Hattie 2009), developing and using data systems to inform and monitor decisions (Lachat and Smith 2005; Supovitz and Klein 2003), creating personalized learning environments (Crosnoe 2011; Robinson et al. 2008), and developing a coherent learning climate conducive to learning (Bryk et al. 2010; Cotton 2000; Sebastian and Allensworth 2012).

Research also suggests that school administrators will achieve success in enhancing instructional quality if they allocate their direct efforts with teachers into nonevaluative channels. Here, four domains receive considerable support from empirical research: providing actionable feedback to teachers (Duke 1990; Hattie 2009; Showers 1985; Joyce and Showers 2002; Walberg 2011), creating professional communities in which teachers share goals, work, and responsibility for student outcomes (Vescio et al. 2008), offering tangible support for the work of teachers (Hattie 2009; Ikemoto et al. 2012), and forging systems in which teachers have the opportunity for ongoing professional learning (Bryk et al. 2010; Joyce and Showers 2002; Robinson et al. 2008; Sebastian and Allensworth 2012).

We approached this review of the literature on teacher evaluation with a long-standing commitment to understanding how school principals achieve positive results for the quality of teaching and learning in schools (e.g., Hallinger and Heck 1998;

Hallinger and Murphy 2012; Heck and Hallinger 2009; Murphy 1990, 2008; Murphy et al. 1987). As such, we began with a sympathetic perspective towards strategies that align with an "instructional leadership" perspective on school improvement. Nonetheless, the efficacy of instructional leadership and school improvement strategies must meet dual criteria of empirical evidence and feasibility.

Based this review of research, we conclude that the latest generation of teacher evaluation models has yet to meet either of these criteria. Consequently, we assert that stronger evidence of impact should be obtained prior to undertaking a major reinvestment of staff time and money into this strategy for school improvement. Though we remain skeptical, we will continue to observe future developments with the understanding that the design and implementation of these strategies remain a work in progress.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics, 25*(1), 93–135.

Atkinson, A., Burgess, S., Croxsonc, B., Gregg, P., Propper, C., Slater, H., & Wilson, D. (2009). Evaluating the impact of performance-related pay for teachers in England. *Labour Economics, 16*(3), 251–261.

Attinello, J., Lare, D., & Waters, F. (2006). The value of teacher portfolios for evaluation and professional growth. *NASSP Bulletin, 90*(2), 132–152.

Baker, E. L., Barton, P., E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R.J., & Shepard, L.A. (2010). *Problems with the use of student test scores to evaluate teachers.* EPI Briefing Paper #278. Washington, DC: Economic Policy Institute

Balfanz, R., & Byrnes, V. (2006). Closing the mathematics achievement gap in high-poverty middle schools: enablers and constraints. *Journal of Education for Students Placed at Risk, 11*(2), 143–159.

Ball, S. (2003). The teacher's soul and the terrors of performativity. *Journal of Education Policy, 18*(2), 215–228.

Barth, R. (1980). *Run school run*. Cambridge. MA: Harvard University Press.

Barth, R. (1986). On sheep and goats and school reform. *Phi Delta Kappan, 68*(4), 293–296.

Bembry, K. L., & Schumacker, R. E. (2002). Establishing the utility of a classroom effectiveness index as a teacher accountability measure. *Journal for Effective Schools, 1*(1), 61–77.

Bidwell, C. E. (1965). The school as a formal organization. In J. G. Marsh (Ed.), *Handbook of organizations* (pp. 972–1022). Chicago: Rand McNally.

Blasé, J., & Kirby, P. (2009). *Bringing out the best in teachers: what effective principals do*. Thousand Oaks: Corwin.

Borman, G. D. (2005). National efforts to bring reform to scale in high-poverty schools: outcomes and implications. *Review of Research in Education, 29*(1), 1–27.

Borman, G., & Kimball, S. (2005). Teacher quality and educational equality: do teachers with higher standards-based evaluation ratings close student achievement gaps? *The Elementary School journal, 106*(1), 3–20.

Bressoux, P., & Bianco, M. (2004). Long-term teacher effects on pupils' learning gains. *Oxford Review of Education, 30*(3), 327–45.

Bridges, E. (1967). Instructional leadership: a concept re-examined. *Journal of Educational Administration, 5*(2), 136–147.

Bridges, E. (1990). *Managing the incompetent teacher* (2nd ed.). Eugene: ERIC Clearinghouse on Educational Management.

Bryk, A. S., Sebring, P. B., & Allensworth, E. (2010). *Organizing schools for improvement: lessons from Chicago*. Chicago: University of Chicago Press.

Callahan, R. E. (1962). *Education and the cult of efficiency.* Chicago: University of Chicago Press.

Camburn, E., Rowan, B., & Taylor, J. E. (2003). Distributed leadership in schools: the case of elementary schools adopting comprehensive school reform models. *Educational Evaluation and Policy Analysis, 25*(4), 347–373.

Castetter, W. B. (1976). *The personnel function in educational administration*. New York: MacMillan.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., Mcpartland, J., Mood, A. M., Weinfeld, F. D., & York, R. T. (1966). *Equality of educational opportunity.* Washington, DC: U.S. Government.

Cotton, K. (2000). *The schooling practices that matter most*. Alexandria: Association for Supervision and Curriculum Development.
Creemers, B., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: a contribution to policy, practice and theory in contemporary schools*. New York: Routledge.
Crosnoe, R. (2011). *Fitting in, standing out: navigating the social challenges of high school to get an education*. Cambridge: Cambridge University Press.
Cuban, L. (1988). *The managerial imperative and the practice of leadership in schools*. Albany: State University of New York Press.
Danielson, C. (2007). *Enhancing professional practice: a framework for teaching* (2nd ed.). Alexandria: Association for Supervision and Curriculum Development.
Darling-Hammond, L., & Youngs, P. (2006). Defining "highly qualified teachers": what does "scientifically-based research" actually tell us? *Educational Researcher, 31*(9), 13–25.
Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8–15.
De Fraine, J., Van Damme, J., & Onghena, P. (2002). Accountability of schools and teachers: what should be taken into account? *European Educational Research Journal, 1*(3), 403–427.
Duke, D. L. (1990). Developing teacher evaluation systems that promote professional growth. *Journal of Personnel Evaluation in Education, 4*, 131–144.
Duke, D. L., Showers, B. K., & Amber, M. (1980). Teacher and shared decision-making: the costs and benefits of involvement. *Educational Administrative Quarterly, 16*(1), 25–35.
Ellett, C., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101–128.
Flores, A. A. (2012). The implementation of a new policy on teacher appraisal in Portugal: how do teachers experience it at school? *Educational Assessment, Evaluation and Accountability, 24*(4), 351–368.
Fullan, M. (2001). *Leading in a culture of change*. San Francisco: Jossey-Bass.
Garet, M. S., & Delany, M. (1988). Students, courses, and stratification. *Sociology of Education, 61*(2), 61–77.
Gates Foundation. (2013). *Measures of effective teaching (MET)*. Downloaded January 14, 2013 from http://www.gatesfoundation.org/united-states/Pages/measures-of-effective-teaching-fact-sheet.aspx.
Glass, G. (2013). *Gates Foundation wastes more money pushing VAM*. Downloaded January 14, 2013 from http://ed2worlds.blogspot.com/2013/01/gates-foundation-wastes-more-money.html.
Gleeson, D., & Husbands, C. (2003). Modernizing schooling through performance management: a critical appraisal. *Journal of Education Policy, 18*(5), 499–511.
Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1). Downloaded on Jan. 3, 2013 from http://educationnext.org/the-mystery-of-good-teaching/.
Goldhaber, D., & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *Review of Economics and Statistics, 89*(1), 134–150.
Gough, D. (2007). Weight of evidence: a framework for the appraisal of the quality and relevance of evidence. *Applied and Practice-based Research, 22*(2), 213–228.
Gray, J., Wilcox, B., Goldstein, H., Hannon, V., Hedger, K., Jesson, D., Rasbash, J., & Sime, N. (1995). *Good school, bad school: evaluating performance and encouraging improvement*. Buckingham: Open University Press.
Grotke, E. (1953). Professional distance and teacher evaluation. *Phi Delta Kappan, 34*(4), 127–130.
Hallinger, P. (2013). A conceptual framework for reviews of research in educational leadership and management. *Journal of Educational Administration, 51*(2), 126–149.
Hallinger, P., & Heck. (1998). Exploring the principal's contribution to school effectiveness: 1980–1995. *School Effectiveness and School Improvement, 9*(2), 157–191.
Hallinger, P., & Murphy, J. F. (2012). Running on empty? Finding the time and capacity to lead learning. *NASSP Bulletin, 97*, 5–21.
Hallinger, P., Ko, J., & Walker, A. (2014). Exploring whole school vs. subject department improvement in Hong Kong secondary schools. *School Effectiveness and School Improvement,* in press.
Hamilton, L. S., Stecher, B. M., Russell, J. L., Marsh, J. A., & Miles, J. (2008). Accountability and teaching practices: school-level actions and teacher responses. In B. Fuller, M. K. Henne, & E. Hannum (Eds.), *Strong states, weak schools: the benefits and dilemmas of centralized accountability*. St. Louis: Emerald (Research in the Sociology of Education, Vol. 16, pp. 31–66).
Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*, 84–117.
Hanushek, E. (2010). *The economic value of higher teacher quality*. Cambridge: National Bureau of Economic Research. Working Paper 16606 http://www.nber.org/papers/w16606.

Hanushek, E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267–71.

Harris, D. N., & Herrington, C. D. (2006). Accountability, standards, and the growing achievement gap: lessons from the past half century. *American Journal of Education, 112*(2), 209–238.

Harvey, L. (2005). A history and critique of quality evaluation in the UK. *Quality Assurance in Education, 13*(4), 263–276.

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

Hawley, W., & Rosenholtz, S. (1984). Good schools: what research says about improving school achievement. *Peabody Journal of Education, 61*, 117–124.

Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal, 46*, 626–658.

Heneman, H., III, & Milanowski, A. T. (2007). *Assessing human resource alignment: the foundation for building total teacher quality improvement*. Madison: Consortium for Policy Research in Education.

Herman, R., & Stringfield, S. (1997). *Ten promising programs for educating all children: evidence of impact*. Arlington: Education Research Service.

Hopkins, D., & Stern, D. (1996). Quality teachers, quality schools: international perspectives and policy implications. *Teaching and Teacher Education, 12*(5), 501–517.

Horng, E. L., Klasik, D., & Loeb, S. (2010). Principal time-use and school effectiveness. National Center for the Analysis of Longitudinal Data in Education research. Retrieved June 1st 2010 from www.stanford.edu/.../Principal%20Time-Use%20Research%20Paper%20(revised).pdf.

Ikemoto, G., Taliaferro, L., & Adams, E. (2012). *Playmakers: how great principals build and lead great teams of teachers*. New York: New Leaders.

Joyce, B., & Showers, B. (2002). *Student achievement through staff development*. Alexandria: Association for Supervision and Curriculum Development.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: looking inside the 'black box' of complex metrics. *Educational Assessment, Evaluation and Accountability, 22*(3), 181–198.

Kimball, S. M., & Milanowski, A. T. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34–70.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education, 79*(4), 54–78.

Koppich, J., & Showalter, C. (2005). *Strategic management of human capital: a cross-case analysis of five districts*. Madison: Strategic Management of Human Capital.

Kyriakides, L., Creemers, B., Antoniou, P., & Demetriou, D. (2009). A synthesis of studies searching for school factors: implications for theory and research. *British Educational Research Journal, 36*(1), 1–24.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk, 10*(3), 333–349.

Latham, G., & Wexley, K. (1981). *Increasing productivity through performance appraisal*. Menlo Park: Addison Wesley.

Lee, M. S., & Hallinger, P. (2012). Exploring the impact of national context on principals' time use: economic development, societal culture, and educational system. *School Effectiveness and School Improvement, 23*(4), 461–482.

Leithwood, K. (2001). School leadership in the context of accountability policies. *International Journal of Leadership in Education, 4*(3), 217–235.

Leithwood, L., & Earl, L. (2000). Educational accountability effects: an international perspective. *Peabody Journal of Education, 75*(4), 1–20.

Leithwood, K., Harris, A., & Strauss, T. (2010). *Leading school turnaround: how successful leaders transform low-performing schools*. San Francisco: Jossey-Bass.

Lewis, A. (2008). *Add it up: using research to improve education and minority students*. Washington, DC: Poverty and Race Research Action Council. Available from http://www.prrac.org/pubs_aiu.pdf.

Liu, S., & Zhao, D. (2013). Teacher evaluation in China: latest trends and future directions. *Educational Assessment, Evaluation and Accountability, 25*(3), 231–250.

Lortie, D. (1975). *School-teacher: a sociological study*. Chicago: University of Chicago Press.

Louis, K. S., Dretzke, B., & Wahlstrom, K. (2010). How does leadership affect student achievement? Results from a national US survey. *School Effectiveness and School Improvement, 21*(3), 315–336.

Loup, K., Garland, J., Ellett, C., & Rugutt, J. (1996). Ten years later: findings from a replication of a study of teacher evaluation practices in our 100 largest districts. *Journal of Personnel Evaluation in Education, 10*(3), 203–26.

Marshall, K. (1996). How I confronted HSPS (hyperactive superficial principal syndrome) and began to deal with the heart of the matter. *Phi Delta Kappan, 76*(5), 336–345.

May, H., & Supovitz, J. A. (2011). The scope of principal efforts to improve instruction. *Educational Administration Quarterly, 47*(2), 332–352.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica: Rand.

Medley, D., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *Journal of Educational Research, 80*(4), 242–267.

Mendro, R. L. (1998). Student achievement and school and teacher accountability. *Journal of Personnel Evaluation in Education, 12*, 257–267.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Meyer, J. W., & Rowan, B. (1975). *Notes on the structure of educational organizations: revised version*. Paper presented at the annual meeting of the American Sociological Association, San Francisco, CA.

Milanowski, A. (2004a). *Relationships among dimension scores of standards-based teacher evaluation systems and the stability of evaluation score/student achievement relationships over time*. Madison: Wisconsin Center for Education Research. CPRE-UW Working Paper Series TC-04-02.

Milanowski, A. (2004b). The relationship between teacher performance evaluation scores and student achievement: evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Milanowski, A.T., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: replication and extensions at three sites*. CPRE-UW Working Paper Series TC-04-01. Madison, WI: Wisconsin Center for Education Research, Consortium for Policy Research in Education

Milanowski, A., Kimball, S., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring school performance and efficiency: implications for practice and research* (pp. 137–161). Washington D.C.: Yearbook of the American Education Finance Association.

Millman, J. (1981). *Handbook of teacher evaluation*. Beverly Hills: Sage.

Millman, J. (1997). *Grading teachers, grading schools; is student achievement a valid evaluation measure?* Thousand Oaks: Corwin Press.

Murphy, J. F. (1990). Principal instructional leadership. In R. S. Lotto & P. W. Thurston (Eds.), *Advances in educational administration: changing perspectives on the school* (Vol. 1, Pt. B, pp. 163–200). Greenwich: JAI.

Murphy, J. F. (1991). *Restructuring schools: capturing and assessing the phenomena*. New York: Teachers College Press.

Murphy, J. F. (2008). *Turning around failing schools: leadership lessons from the organizational sciences*. Thousand Oaks: Corwin Press.

Murphy, J., Hallinger, P., Lotto, L., & Miller, S. (1987). Barriers to implementing the instructional leadership role. *The Canadian Administrator, 27*(3), 1–9.

Murphy, J. F., Hallinger, P., & Heck, R. H. (2013). Leading via teacher evaluation: the case of missing clothes? *Educational Researcher, 42*(6), 349–354.

Musella, D. (1970). Improving teacher evaluation. *Journal of Teacher Education, 21*(1), 15–21.

Odden, A. (2004). Lessons learned about standards-based teacher evaluation systems. *Peabody Journal of Education, 79*(4), 126–137.

Odden, A., & Wallace, M. (2008). *How to achieve world class teacher compensation*. Indianapolis: Freeload.

Popham, W. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education, 1*(3), 269–273.

Purkey, S., & Smith, M. (1983). Effective schools: a review. *The Elementary School Journal, 83*(4), 426–452.

Range, B., Scherz, S., Holt, C., & Young, S. (2011). Supervision and evaluation: the Wyoming perspective. *Educational Assessment, Evaluation and Accountability, 23*(3), 243–265.

Reyes, P., Scribner, J., & Scribner, A. (1999). *Lessons from high-performing Hispanic schools: creating learning communities*. New York: Teachers College.

Reynolds, D., Teddlie, C., Hopkins, D., & Stringfield, S. (2000). Linking school effectiveness and school improvement. In C. Teddlie & D. Reynolds (Eds.), *The international handbook of school effectiveness research* (pp. 206–231). London: Falmer.

Reynolds, D., Muijs, D., & Treharne, D. (2003). Teacher evaluation and teacher effectiveness in the United Kingdom. *Journal of Personnel Evaluation in Education, 17*(1), 83–100.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2000). *Teachers, schools, and academic achievement*. Cambridge: National Bureau of Economic Research. NBER Working Paper # W6691.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*, 417–458.
Robinson, V. M. J., & Timperly, H. (2007). The leadership of the improvement of teaching and learning. *Australian Journal of Education, 51*(3), 247–262.
Robinson, V. M. J., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: an analysis of the differential effects of leadership types. *Educational Administration Quarterly, 44*(5), 635–674.
Rockoff, J. (2004). The impact of individual teachers on student achievement: evidence from panel data. *The American Economic Review, 94*(2), 247–252.
Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2), 261–66.
Rosenholtz, S. J. (1991). *Teachers' workplace: the social organization of schools*. New York: Teachers College Press.
Rothstein, J. (2009). *Student sorting and bias in value added estimation: selection on observables and unobservables*. Cambridge: National Bureau of Economic Research. Working Paper, 14666.
Rowan, B. R., Correnti, R., & Miller, R. J. (2002). What large-scale survey research tells us about teacher effects on student achievement: insights from the prospects study of elementary schools. *Teachers College Record, 104*, 1525–1567.
Sanders, W., & Horn, S. (1994). The Tennessee value-added assessment system (TVASS). Mixed-methods model methodology in educational assessment. *Journal of Personnel Evaluation in Education, 8*, 299–311.
Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
Sanders, W., Ashton, J., & Wright, S. (2005). *Comparison of the effects of NBPTS-certified teachers with other teachers on the rate of student academic progress*. Washington, DC: U.S. Department of Education and National Science Foundation.
Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
Sebastian, J., & Allensworth, E. (2012). The influence of principal leadership on classroom instruction and student learning: a study of mediated pathways to learning. *Educational Administration Quarterly, 48*(4), 626–663.
Showers, B. (1985). Teachers coaching teachers. *Educational Leadership, 42*(7), 43–49.
Skedsmo, G. (2011). Formulation and realisation of evaluation policy: inconcistencies and problematic issues. *Educational Assessment, Evaluation and Accountability, 23*(1), 5–20.
Slavin, R., Karweit, N., & Madden, N. (1989). *Effective programs for students at risk*. Boston: Allyn and Bacon.
Spillane, J. P., Camburn, E., & Pareja, A. (2009). School principals at work: a distributed perspective. In K. Leithwood, B. Mascall, & T. Strauss (Eds.), *Distributed leadership according to the evidence* (pp. 87–110). London: Routledge.
Stiggins, R., & Duke, D. (1988). *The case for commitment to teacher growth: research on teacher evaluation*. Albany: SUNY Press.
Supovitz, J. A., & Klein, V. (2003). *Mapping a course for improved student learning: how innovative schools systematically use student performance data to guide improvement*. Philadelphia: Consortium for Policy Research in Education.
Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. New York: Falmer Press.
Thomas, S. (2001). Dimensions of secondary school effectiveness: comparative analyses across regions. *School Effectiveness and School Improvement, 12*(3), 285–322.
Toch, T., & Rothman, R. (2008). *Rush to judgment: Teacher evaluation in public education*. Available at: www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf. Accessed 14 Jul 2013.
Tyack, D. B. (1974). *One best system*. Cambridge: Harvard University Press.
Vescio, V., Ross, D., & Adams, A. (2008). A review of research on the impact of professional learning communities on teaching practice and student learning. *Teaching and Teacher Education, 24*(1), 80–91.
Walberg, H. (2011). *Improving student learning: action principles for families, classrooms, schools, districts, and states*. Charlotte: Information Age.
Walker, A. D., & Ko, J. (2011). Principal leadership in an era of accountability: a perspective from the Hong Kong context. *School Leadership & Management, 31*(4), 369–392.
Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools; is student achievement a valid evaluation measure?* (pp. 81–99). Thousand Oaks: Corwin.
Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21*, 1–19.

White, B. (2004). *The relationship between teacher evaluation scores and student achievement: evidence from Coventry, RI*. CPRE-UW Working Paper Series TC-04-04. Madison, WI: University of Wisconsin-Madison, Wisconsin Center for Education Research, Consortium for Policy Research in Education, San Diego, CA.

Wilson, M., Hallman, P. J., Pecheone, R., & Moss, P. (2014). Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's Beginning Educator Support and Training Program. *Education Evaluation and Policy Analysis*. in press.

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

Wright, S., Horn, S., & Sanders, P. (1997). Classroom context effects on student achievement: implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57–67.