2013

A Review of Literature on Teaching Evaluation

Martin Reardon
Virginia Commonwealth University, rmreardon@vcu.edu

A Review of Literature on Teaching Evaluation

Martin Reardon

Metropolitan Educational Research Consortium

MERC

Metropolitan Education Research Consortium

MERC Membership

Chesterfield County Public Schools

Colonial Heights City Schools

Goochland County Public Schools

Hanover County Public Schools

Henrico County Public Schools

Hopewell City Public Schools

Powhatan County Public Schools

Richmond City Public Schools

Virginia Commonwealth University

Background

Virginia Commonweatlh Univeristy and the school divisions of Chesterfield, Colonial Heights, Hanover, Henrico, Hopewell and Richmond established the Metropolitan Educational Research Consortium (MERC) on August 29, 2991. The founding members created MERC to provide timely information to help resolve education problems identified by practicing professional educations. MERCC membership is open to all metropolitan-type school divisions. It currently provides services to over 12,000 teachers and 152,000 students. MERC has based funding from its membership. Its study teams are composed of university investigators and practitioners from the membership.

MERC is organized to serve the interests of its members by providing tangible material support to enhance the practice of educational leadership and the improvement of teaching and learning in metropolitan educational settings. MERC's research and development agenda is built around four goals:

- To improve educational decision-making through joing development of practice-driven research questions, design and dissemination,
- To anticipate important educational issues and provide leadership in school improvement
- To identify proven strategies for resolving instruction, management, policy and planning issues facing public education, and
- To enhance the dissemination of effective school practices.

In addition to conducting research as described above, MERC conducts technical and educational seminars, program evaluations, and publishes reports and brief on a variety of educational issues.

A Review of Literature on Teaching Evaluation

Table of Contents

Introduction p4

The Imperative for Change p4

Section 1: Leading up to NCLB p6

Section II: After NCLB p7

Section III: More Recent Perspectives P8

Section IV: At the Cutting Edge p10

Conclusion p13

References p13

Introduction: A Review of Literature on Teaching Evaluation

This review of the literature on teacher evaluation was developed at the invitation of the Policy and Planning Council of the Metropolitan Educational Research Consortium (MERC) in the context of the current focus on teacher evaluation in Virginia. Consequently, in this document, high priority was accorded to the areas of focus formulated by the *Virginia Standards for the Professional Practice of Teachers* (*Standards*, Virginia Department of Education, 2011), the *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers* (*Guidelines*, Virginia Department of Education, 2011) and *The Research Base for the Uniform Performance Standards for Teachers* (*Research Base*, Virginia Department of Education, 2011). As the *Commonwealth of Virginia Department of Education, Superintendents Memo #136-11* explained,

> the *Virginia Standards for the Professional Practice of Teachers* define what teachers should know and be able to do, and they establish a foundation upon which all aspects of teacher development from teacher education to induction and ongoing professional development can be aligned. The revised *Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers* incorporate these teaching standards" (p. 1).

Part 1 of the *Research Base* (2011) repeats the performance standards from Part 5 of the *Guidelines* (2011). Part 2 opens with an acknowledgment of a "high degree of alignment" (p. 12) between the seven uniform performance standards for teachers in Virginia and the standards promulgated by both the Interstate New Teacher Assessment and Support Consortium (INTASC) and the National Board for Professional Teacher Standards (NBPTS)—except that neither INTASC nor NBPTS include measures of student academic progress. The remainder of Part 2 provides insight into the literature that supports each standard. This insight is helpful, and perhaps all that is necessary in such a document. However, the discussion of the nine individual literature references in the one and one-half pages supporting Student Academic Progress (Standard 7) does not constitute a strong defense of the addition of a standard that is absent from both the INTASC and NBPTS standards.

The members of MERC's Policy and Planning Council are well aware of the challenges that school divisions are facing as they conscientiously endeavor to honor the stipulations of the *Guidelines* (2011) in a loosely coupled system. In keeping with the context out of which the request for this literature review arose, a sense of the literature invoked in the *Research Base* (2011) has been taken for granted. The focus of this document, then, is on literature—predominantly recent literature—that addresses issues associated with the implementation of teacher evaluation. This emphasis on implementation is in keeping with MERC's mission to engage in research that school divisions can use.

The Imperative for Change

In the first section of *Accelerating the Agenda* (2008), the National Governors Association, the National Conference of State Legislatures, the National Association of State Boards of Education, and the Council of Chief State School Officers joined to assert that "improvement in student learning can dramatically boost economic growth," that "readiness for college and career remains more relevant today than ever before," that "U.S. students are exiting high school with weaker skills than their counterparts of 20 years ago," and that "it is up to states to lead this charge for college- and career-readiness" (pp. 3-5). These beliefs have been supported by the national policies that have been in place throughout the current U.S. Presidency.

According to the Executive Director of Research and Strategic Planning at the Virginia Department of Education, the purpose of Virginia's focus on teacher evaluation is to improve student achievement with a particular focus on high-poverty and/or persistently low-performing schools (Jonas, 2011, personal communication). According to Jonas (2011, personal communication), in terms of the *Standards* (2011), the rationale underpinning teacher evaluation is that the performance of students is likely to show strong and measurable learning gains (the seventh standard) if students are taught by teachers whose practice exemplifies the first six standards (professional knowledge, instructional planning, instructional delivery, assessment of and for learning, learning environment, and professionalism). The evaluation of teachers using "multiple ways over time" (*Guidelines*, 2011, p. 41), the rationale continues, will reliably measure teacher's ability to add value to students' learning and drive professional development, or, in extreme cases, provide grounds for dismissal.

This entire rationale is plausible at the policy level, but there are inherent subtleties in its implementation. For example, there is evidence for a decrease in returns attributable to teachers' learning "on-the-job" up to about the third year of teaching, with a plateau thereafter (Henry, Bastian, & Fortner, 2011). A body of research confirms that students taught by teachers who are in the second year of teaching post larger average achievement gains than similar students in similar schools who are taught by beginning teachers (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006; Clotfelter, Ladd, & Vigdor, 2007, 2010; Goldhaber, 2007; Kane, Rockhoff, & Staiger, 2008; Rivkin, Hanushek, & Kain, 2005). In the same way, the students of teachers in their third year of teaching post analogous but less large average achievement gains compared to the students of their second-years peers (Boyd et al., 2006; Kane et al., 2008; Staiger & Rockoff, 2010).

However, the magnitude of the statistically significant gains that have been reported in these years during which there is consensus that the greatest changes take place translate into educational differences that would be arguably invisible to a supervisor. For example, Henry, Bastian, and Fortner (2011) cite the first-year, second-year, and third-year teacher deficits in their North Carolina study as -0.043, -0.010, and -0.001 standard deviation units respectively. The question for the supervisor is how deficits of this magnitude can be detected in the walk-through conducted last week, in the third hour formal observation of a teacher scheduled for the first Tuesday in March, in the returns from a teacher's student survey, in a teacher's student goal setting report, or in the penultimate section of a teachers' annual review portfolio—to list just a few of the teacher evaluation approaches explored in the *Guidelines* (2011). Further challenging the nexus between performance on the first six standards and student academic performance gains, Henry et al. concluded that "prior research has overestimated returns to experience, as both teacher on-the-job training and the differential attrition of less effective teachers contribute to the apparent gains in average first - and second-year teachers' effectiveness" (p. 278).

This rest of this document is divided into four non-exclusive, conceptually related sections. Section 1 provides a thumbnail sketch of the history of teacher evaluation leading up to the *No Child Left Behind Act of 2001* (NCLB, 2002). In the five or six years prior to the NCLB milestone, an informal consensus emerged about the elements of best practice in teacher evaluation, and how these elements could be incorporated into viable models. These models remain viable, and 10 of them were incorporated to varying degrees in the models proposed in the *Guidelines* (2011). Section 2 provides an overview of aspects of the legacy consensus in the light of the post-NCLB literature. Section 3 sharpens the focus of the preceding discussion, looking specifically at literature from the past five years. Section 4 suggests

where the cutting edge of contemporary approaches may lie, and suggests that holistic, systemic approaches may supplement the individual teacher evaluations and motivate system-wide improvements.

Section I: Leading up to NCLB

> *The estimated difference in annual achievement growth between having a good and having a bad teacher can be more than one grade-level equivalent in test performance.*
>
> Eric A. Hanushek, Hoover Institution, Stanford University (1992, p. 107)

In a comprehensive overview of teacher evaluation up to the mid-1990s, Shinkfield and Stufflebeam (1995) drew attention at the outset to the fact that formal evaluation of teachers is a relatively recent phenomenon. They described the practice as "virtually unknown until the turn of the 20th century" (p. 9), but even then not gaining momentum until the 1970s. They identified *A Nation at Risk* (1984) as a significant catalyst, but they critiqued the widespread adherence to practices of teacher evaluation as "motivated as much by the enactment of state legislative requirement as the desire to improve the professional status of teachers" (p. 9). It would indeed be the exceptional school division that did not highly value the improvement of the professional status of its teachers, but it is also unlikely that as much effort would be being expended to implement defensible teacher evaluation processes in the absence of legislative requirements.

Shinkfield and Stufflebeam (1995) also listed three challenges for the future of teacher evaluation that have proved to be prescient (a) the association of teacher evaluation with merit pay, (b) the decision about who should be involved in the evaluation process, and (c) the untangling of the differential impact of teaching styles, contexts, and social environments on student achievement. Further, they pointed out the inherent difficulty of combining the functions of formative (developmentally oriented) and summative (oriented to the meeting of consumers' needs) evaluations (Scriven, 1967), and mused that "history may record that (the formative and summative functions) are incompatible unless they are controlled and administered separately" (p. 31).

Table 1

Fifteen Models of Teacher Evaluation and Contemporary Instances

Distinctive Feature of Model/	Comment
Traditional Impressionistic	Judgments are made on the basis of the observers' experience and educational
Clinical Supervision	Related to Madeleine Hunter's approach. See Glickman, Gordon, and Ross-
Research-based Checklist *Guidelines*, pp. 16-18, & 20-21	The basis of many state-mandated evaluation instruments, and two prominent commercial systems: Danielson's *Framework for Teaching*, and Marzano's *Teacher Evaluation Model*.
High Inference Judgments	Trained evaluators ensure reliable judgments. See Pianta, La Paro, and Hamre (2008); *Classroom Assessment Scoring System* (CLASS).
Interviewing	Wide range of uses on an ongoing basis.
Tests of Professional Knowledge	National Board of Professional Teaching Standards (NBPTS); Interstate New
Management by Objectives	Develop goals and indicators, and review regularly. See
Job Analysis	Related to competency-based teacher evaluation. Artifact, portfolios, document
Duties-based Evaluation *Guidelines*, pp. 28-34	Based on Scriven's (1988, 1994) studies of what teachers legally can be expected to do.
Theory-based Approach	Based on evaluative criteria derived from theory that, for example, links student achievement to specific teaching practices. See Danielson's *Framework for Teaching*; Slavin and colleagues' *Success for All*.
Student Achievement *Guidelines*, pp. 42-45	Based on value-added measures of student achievement at specific times. See Betebenner (2009); Blank (2010). Alternatively, measures "grounded in validated, quantitative, objective measures, using tools already available in the school" (*Guidelines*, 2011, p. 42).
Client Ratings *Guidelines*, pp. 22-27	Student and (potentially) parent ratings. See Stronge and Ostrander (1997).
Peer Evaluation	For example, a peer sitting on a review panel for a colleague. A component of three implementations of teacher evaluation in Shinkfield and Stufflebeam (1995).
Self-evaluation	Teachers evaluate their own performance on specific teaching criteria. See Na-
General Criteria	See *Virginia Standards for the Professional Practice of Teachers* (2011, pp. 3-5).

Note: Edited and updated from *Teacher Evaluation: Guide to Effective Practice*, by A. J. Shinkfield and D. Stufflebeam, 1995, pp. 175-176.

[1]This abbreviation refers to *Guidelines for Uniform Performance Standards and Evaluation Criteria* (2011).

Shinkfield and Stufflebeam (1995) referred to earlier work in laying out 15 models of "ways to evaluate teachers that implicitly define good teaching" (p. 175). These models (listed with contemporary references in Table 1) cover the gamut of teacher evaluation, and provide a useful reference point. They are models in the sense that they focus on the main concepts of teacher evaluation that the authors of various approaches aligned with these models believe are important. In the first column of Table 1, the distinctive feature of model is followed by a reference to the *Guidelines* (2011), where appropriate. The comments in Table 1 are offered as brief explanations, and to suggest some noteworthy contemporary instances. The first four models rely on classroom observation.

Five years after Shinkfield and Stufflebeam (1995) and shortly before the NCLB (2002) landmark, Peterson (2000) made the point that using student achievement as evidence of teacher quality made sense—especially to noneducators—and that omitting such a yardstick from a measure of teacher quality was "not credible to many audiences" (p. 136). Of course, in NCLB, student achievement was enshrined as the yardstick, regardless of sense or credibility. However, Peterson went on to highlight three problems associated with the use of student achievement as a yardstick of teacher quality that are are still relevant today: (a) teacher quality and student achievement are logically but only indirectly associated (e.g., student disinterest can lessen the impact of the most brilliant teaching), (b) the collection of defensible data on teacher effects on student achievement is technically difficult (e.g., questionable validity and reliability of measures), and (c) such a yardstick distorts the education system (e.g., potentially downplaying important non-tested aspects of schooling like the ability of students to collaborate with others or think critically). As Peterson wryly summarized, student achievement is "a most compelling source for teacher evaluation, if only evaluators can defensibly get it for the teacher under review" (p. 136). Assuredly, this is still the crux of the matter.

Section II: After NCLB

The teacher evaluation process gained considerable traction as one component of the imperative to leave no child behind educationally. NCLB (2002) was intentionally designed to usher in a new era in public education. The hallmark of the new era was that teachers and principals would be held accountable for student achievement. In order for teachers and principals to be accountable, they had to be evaluated. The rationale of NCLB was that better teachers teach students whose achievements meet the standards as defined by the state. As Harris (2011) asserted, teachers and principals should only be held accountable for what they can control, and some highly skilled teachers may be judged harshly if measured by student achievement on state-mandated tests because of factors beyond their control like class size (Glass, Cahen, Smith, & Filby, 1982; Mosteller, 1995), teacher qualifications (Ferguson, 1991), or school size (Monk & Haller, 1993). Some students are motivated by excellent teachers in non-academic areas in which achievement is not measured on state tests, and some involved in education (e.g., excellent librarians and superb counselors) have an even more indirect effect on student achievement in state-mandated tests (Amrein-Beardsley & Collins, 2012).

The simple global concept that teacher and principal accountability will produce improved academic outcomes for students became even more complex as, over time, the national policy was customized at the state level, adapted at division level, and implemented at the individual school level. Principals had been evaluating the teachers in their schools well before NCLB (2002), but the conduct of those evaluations had been seen of part of the professional duty of the principal—as an indication that the principal took seriously his or her

instructional leadership role. NCLB's insistence on accountability initiated an evaluative approach that cast principals in a different role. A year into the NCLB era, the National Commission on Teaching and America's Future (NCTAF, 2003), in considering how to build professionally rewarding career paths in teaching (and thereby stem the hemorrhaging of teachers from the profession), cited "countless studies," to support the engagement of teachers in the analysis of their own practice as a major factor. A second major factor, NCTAF suggested, was the provision of opportunities for teachers to observe and be observed by experts "with strong feedback" (p. 28). NCTAF viewed these two factors as foundational elements in supporting sustained growth of teachers, and neither of these factors directly implicated the principal, but under NCLB (2002), the concept of "strong feedback" assumed ominous overtones. If the teacher evaluation process left in place teachers whose students' performance on achievement tests did not meet standards, then the principal was accountable and his or her position was in jeopardy.

Two years after the NCTAF (2002) report, Peterson (2004) published an overview of research on teacher evaluation for the National Association of Secondary School Principals. Peterson reviewed the research literature, critiqued principal observation as the sole basis for evaluating teachers, and argued for the use of multiple data sources. However, a compelling section of Peterson's paper discussed the sociology of teacher evaluation. Under this heading, Peterson discussed the imperative to consider the "sociological balance" (p. 73) of the school division in evaluating teachers. He pointed out that collecting data by means of peer review of materials or client surveys (Stronge & Ostrander, 1997) affects the status of teachers by replacing "casual hearsay" (p. 73) as the basis of esteem with publicly accessible information.

To return to the use of multiple data sources, in 2006, Peterson contributed a chapter on the use of multiple data sources in teacher evaluation systems to the second edition of a text edited by Stronge (2006). Peterson contended that multiple sources are relevant (a) because the complexity of teaching renders the reliance on a single source problematic (the variety of contexts in which a teacher teaches are unlikely to be adequately assessed by a single source of information), and (b) because no single source of data is "valid or feasible for each and every teacher in a school district" (p. 215). The chapters included in Stronge's (2006) text provide a list of multiple sources of data:

- Classroom-based assessment of teaching & learning
- Client surveys
- Student achievement
- Portfolios
- Teacher self-evaluation

Section III: More Recent Perspectives

> *Even in pilot projects, (value-added methodology) estimates of teacher effectiveness that are based on data for a single class of students should not be used to make operational decisions because such estimates are far too unstable to be considered fair or reliable.*
>
> Letter Report to the U.S. DOE on the Race to the Top Fund,
> Board on Testing and Assessment,
> National Academy of Sciences (p. 10)

The one source of data that was missing in Stronge's (2006) text that has received a great deal of attention in recent years is the use student achievement test scores to evaluate teachers. The most visible of the approaches towards the end of the last century to show the impact on student achievement scores of effective and ineffective teachers were associated with the work of Sanders and his colleagues in Tennessee (for example, Sanders & Rivers, 1996; see Reardon, 2011).

The early and subsequent work in this field has been called into question by those who are skeptical of both the validity and reliability of the results as applied to

individual teachers. For example, in recent years, the Board on Testing and Assessment (BOTA, 2009) applauded the *Race to the Top Fund* emphasis on the removal of legislative obstacles in any funded state to the creation of data systems that link students and teachers—a necessary prerequisite for the use of student achievement data as measures of teacher quality. However, the caveat from BOTA at the start of this section is only one of their cautions. BOTA's concerns grew from supporting literature, and similar reservations have been expressed subsequently (Schochet & Chiang, 2010). For example, Baker et al. (2010) cited a study that found that "students' fifth grade teachers were good predictors of their *fourth* grade test scores" (p. 2). Even in the absence of an outcome that is as startling as that discussed by Baker et al., Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2012) suggested that when the specific class or grade-level assignment was a stronger predictor of the value-added rating than the teacher, then reference to a teacher effect was problematic.

Reardon (2011) set student percentile growth modeling (Virginia's version of value-added growth modeling) in the context of the value added literature. In summary, Reardon agreed that Betebenner's (2009) intentional retreat from the accuracy of individual student scores to percentile ranking avoided some of the more obvious pitfalls of score-based value-added approaches.

However, score-based value-added measures do have a role in research (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). In an edited volume that gathered the papers from a 2008 conference on teacher effects hosted by the University of Notre Dame's Center for Research on Educational Opportunity, Kelly (2012) cited data from Milwaukee and surmised that the primary reason that children in schools serving neighborhoods with high concentrations of poverty score lower on achievement tests is the influence of high-poverty home and neighborhood environments—not low teacher quality. Kelly's surmise is supported by the findings of a much earlier longitudinal study by Entwistle, Alexander, and Olson (1997) that noted nearly identical achievement growth across high- and low-SES schools, except in summer when SES-related educational opportunities varied widely. The Entwistle et al. finding was replicated by Downey, von Hippel, and Hughes (2008), and highlighted by Darling-Hammond et al. (2012). The point is that a solid body of evidence fails to support the attribution of the most "obvious" student achievement discrepancies—those across schools—to variance in teacher quality.

When it comes to student achievement discrepancies within schools, however, Konstantopoulos (2012) asserted that student achievement varies considerably depending on which teacher the student is assigned. Konstantopoulos cited research from the past 30 years to support his assertion, although three of the four works he cited are much more recent than that time range suggests (Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005; Rowan, Correnti, & Miller, 2002).

If student achievement varies considerably depending on the teacher, then it is reasonable to suggest that the characteristics of the teacher may be at least partially responsible. There are many list of characteristics of effective teachers, but one comprehensive list is that provided by Darling-Hammond and Bransford (2005), who suggested that effective teachers

- Understand subject matter deeply and flexibly;
- Connect what is to be learned to students' prior knowledge and experience;
- Create effective scaffolds and supports for learning;
- Use instructional strategies that help students draw connections, apply what they're learning, practice new skills, and monitor their own learning;

- Assess student learning continuously and adapt teaching to student needs;
- Provide clear standards, constant feedback, and opportunities for revising work; and
- Develop and effectively manage a collaborative classroom in which all students have membership.

There is great scope for evaluating teacher performance on any one of the above seven characteristics through observation, or any of the other approaches that honor the context of teaching. However, a strong incentive for turning to more quantitative approaches has been the noteworthy failure of conventional, qualitatively oriented evaluation processes to promote improvement in student achievement.

There are many instances in the literature where researchers have critiqued the conventional processes. For example, von Frank (2011) focused on a district where 99% of the 12,000 teachers were rated satisfactory or outstanding and "nearly half of high school teachers received perfect scores" (p 32). This was a high-performing district, but as von Frank commented "many in the district agreed the evaluations must be misleading" (p. 32). Weisberg, Sexton, Mulhearn, and Keeling (2009) concluded from a twelve-district, four-state study that "a teacher's effectiveness—the most important factor for schools in improving student achievement—is not measured, recorded, or used to inform decision-making in any meaningful way" (p. 1). The New Teacher Project (2007) found that "87% of (Chicago's) 600 schools, including 69 schools that the city declared to be failing, did not issue a single 'unsatisfactory' teacher rating between 2003 and 2006" (Toch, 2008, p. 32).

Donaldson (2010) referred to a situation in which teacher evaluations were full of valentines—"vague, meaningless praise—largely devoid of constructive criticism or concrete feedback" (p. 54). As suggested above, it would be expected that teachers within any school would vary across a list of characteristics of effectiveness as impressive as that provided by Darling-Hammond and Bransford (2005). Donaldson cited her earlier research in which she showed that "on the whole, teacher evaluation has not substantially improved instruction" (p. 54). Danielson (2002) insisted that the two fundamental purposes of teacher evaluation were "quality assurance and professional learning" (p. 64), so clearly something is amiss if improved instruction was not evident as a result of teacher evaluation. Donaldson proposed that one of the reasons for the ineffectiveness of teacher evaluations in promoting improved instruction were the valentines that instructional leaders awarded to teachers.

Section IV: At the Cutting Edge

> *If you select the right measures, you can provide teachers with an honest assessment of where they stand in their practice that, hopefully, will serve as their launching point for their development.*
>
> Thomas J. Kane, Harvard Graduate School of Education, MET Director

Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein (2012) referred to "a growing consensus that evidence of teacher contributions to student learning should be part of teacher evaluation systems, along with evidence about the quality of teacher practices" (p. 8). Darling-Hammond et al. went on to recommend the use of professional standards to evaluate teachers because evaluations based on the professional standards (for example, INTASC standards) "produce ratings that are much more stable than value-added measures" (p. 13). Darling-Hammond et al. highlighted the work of the Measures of Effective Teaching (MET) project in this area.

The MET project was a three-year endeavor funded by the Bill & Melinda Gates Foundation to "identify great

teaching and empower teachers to help students to succeed" (http://www.metproject.org/more.php). The culminating findings of MET were released in January, 2013, when Kane made the comment quoted at the start of this Section in an interview for *Education Week* (Sawchuk, 2013). Taking into account the elegance of the research design, the participation of 3,000 teachers from seven widely dispersed and diverse school districts, the credentials of the principal investigators, and the caliber of the research partners, MET represents a definitive snapshot of the state of the art in terms of teacher evaluation at this time.

As indicated on the MET website, MET researchers collected data in five areas related to effective teaching (http://www.metproject.org/more/components.php). The understanding supporting the research in these five areas emerged from an interpretation of Rivkin, Hanushek, and Kain (2005) to the effect that "the teacher has more impact on student learning than any other factor controlled by school systems, including class size, school size and the quality of after-school program, or even which school a student is attending" (MET, 2010a, 2010b, 2010c, 2010d, 2010e, p. 1 in each document). Short explanations of the five areas are as follows:

1. Student achievement gains on state standardized tests and supplemental tests

MET calculated two value-added estimates—one based on the particular state's assessment scores, and another based on supplemental tests, including the ACT QualityCore series for Algebra I, English 9, and Biology, the Balanced Assessment in Mathematics for grades 4 through 8, and the Stanford 9 Open-ended Reading Assessment for grades 4 through 8 (MET, 2010c)

2. Classroom observations and teacher reflections

MET investigated teachers' classroom instruction styles. Videos of classroom teaching accompanied by written commentary, supporting/contextual material, and the personal reflections of the teacher were analyzed by trained and supervised analysts (MET, 2010a).

3. Teachers' pedagogical content knowledge

As the *Research Base* (2011) also notes, Shulman (1986) discussed pedagogical content knowledge as the second of three categories of content knowledge—a knowledge that "goes beyond knowledge of subject matter per se to the dimension…of content knowledge that embodies the aspects of content most germane to its teachability" (p. 9). Prior to Shulman's article, Byrne (1983) had stressed also the importance of both content knowledge and how to teach that content. Teachers involved in the MET project took assessments to evaluate their pedagogical content knowledge in relation to, for example, their ability to evaluate student understanding and diagnose common student errors (MET, 2010b).

4. Student perceptions of the classroom instructional environment

MET utilized surveys developed over the past decade by Cambridge Education to "assess whether or not students agree with a variety of statements designed to measure seven teaching practices" (MET, 2010d, p. 1) that have been shown to be related to higher average student achievement scores.

5. Teachers' perceptions of working conditions and support at their schools

MET (2010e) cited the work of Ladd (2009) who, from a North Carolina study, concluded that "working conditions variables account for 10 to 15 percent of the explained variation in math and reading scores across schools, after controlling for individual and school level characteristics of schools" (p. 37).

The essence of the findings from analyzing the "unprecedented data collected by the MET project over the past three years" (Measures of Effective Teaching, 2013, p. 21) include:

- Student perception surveys and classroom observations can provide meaningful feedback to teachers,
- Training and certification of observers, and averaging the observations of multiple lessons by different observers can boost the reliability of observations,
- Student learning gains can help identify groups of teachers who are helping students learn more (MET uses the term "cause" to describe the impact of effective teachers because of the random assignment of students to classrooms, however the random assignment was not strongly maintained), and
- A balanced approach in which student achievement gains account for between 33% and 50% of the evaluation leads to more consistent teacher evaluations.

The findings from MET demand attention for the reasons highlighted above. The arbitrary allotment 40% of a teacher's evaluation to measures of student achievement stipulated in the *Guidelines* (2011) is supported by the MET findings. The classroom observations conducted in the context of MET were focused on Domains 2 and 3 of Danielson's *Framework for Teaching* (originally published in 1996). The use of Danielson's *Framework* has been greatly facilitated by a wide range of support functions (see http://www.danielsongroup.org/article.aspx?page=frameworkforteaching), These support functions have encouraged large-scale implementation of the *Framework* (for example, it is the default approach to teacher evaluation in Maryland). A strong alternative approach is that offered by Marzano's (2013) *Teacher Evaluation Model*. These two are arguably the most prominent current approaches to digitally enabled observations as part of teacher evaluation, and both offer extensive support for the task. Instructional leaders can activate remote video recorders in teachers rooms, share videos of exemplary teaching episodes from extensive pre-coded video libraries, and communicate with teachers in a social network environment. Although both Danielson and Marzano make strong claims for the scientific basis for their observational approaches, McCutcheon's caveat that an observation tool limits what is perceived—maybe to the point where the observation may "misrender the classroom" (p. 9) is well made.

The MET findings regarding the training and utilization of observers point to the need to invest in developing expertise among a number of observers, who will then conduct multiple observations. If approached in the most obvious way (intensive professional development of large numbers of observers), the MET finding in this regard seems quite impractical in the absence of the financial backing that MET enjoyed. However the inability to mimic the research conditions should not discourage close approximations. One such local approximation will serve to conclude this section.

In an endeavor to maximize the return for the investment of resources in raising student achievement, a current research project is establishing a blueprint that leverages the sociology of teacher evaluation (Peterson, 2004) and peer review in a systemic approach to instructional improvement. The blueprint touches many of the bases of improvement that have been addressed in this review of literature. The blueprint utilizes within-school instructional rounds (City, Elmore, Fiarman, & Teitel, 2009) staffed by teachers to take the academic pulse of the school. In brief overview, the learning from the instructional rounds visits constitutes the input into small professional learning community structures, out of

which lesson study visits are motivated and debriefed. According to the blueprint, this entire process continues throughout the year, with two instructional rounds visits per semester. In a test-bed middle school, participants in this blueprint trial have observed noticeable positive changes over the course of 18 months in both student engagement and lesson structure. Students' perspectives on hope and engagement are being gathered by means of an externally administered survey. Whether through this blueprint or other alternative approaches that build on the literature, perhaps the gathering of data from multiple sources through the use of multiple approaches can enable a reasonable approximation to best practice.

Conclusion

Teacher evaluation remains a signature task for instructional leadership in schools. The *Guidelines* (2011) constrain a range of percentages of the evaluation of various teachers' performance, but, to a large degree, alternative approaches for the remaining percentage for some teachers, and the total percentage for the majority of teachers, is at the discretion of the school division. Thus, for the majority of teachers, the evaluation model that is implemented is constrained only by the policy adopted in the school division. It is this context that an overview of the extant literature is most relevant. The development of systemic approaches which approximate best-practice, well-funded research, and seek to optimize the return from the considerable resources committed to teacher evaluation may generate a collaborative professional culture that manifests and facilitates the ongoing refinement of effective teaching.

References

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS EVAAS) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12), 1-31.

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

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

Blank, R. K. (2010). *State growth models for school accountability: Progress on developing and reporting measures of student growth*. Washington, DC: Council of Chief State School Officers.

Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education, National Academy of Sciences. (2009, October). Letter report to the U.D. Department of Education on the Race to the Top Fund. Retrieved from http://books.nap.edu/openbook.php?record_id=12780&page=11

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1*, 176-216.

Byrne, C. J. (1983). *Teacher knowledge and teacher effectiveness: A literature review, theoretical analysis and discussion of research strategy*. Paper presented at the meeting of the Northwestern Educational Research Association, Ellenville, NY.

City, E. A., Elmore, R. F., Fiarman, S. E., & Teitel, L. (2009). *Instructional rounds in education: A network approach to improving teaching and learning*. Cambridge, MA: Harvard Education Press.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of*

Education Review, 26, 673-682.

Clotfelter, C., Ladd, H., & Vigdor, J. (2010). Teacher credentials and student achievement in high school: A cross-subject analysis with student fixed effects. *Journal of Human Relations, 45*, 655-681.

Commonwealth of Virginia Department of Education, Superintendents Memo #136-11. (2011, May 13). *Revised "Guidelines for Uniform Performance Standards and Evaluation Criteria for Teachers" and "Virginia Standards for the Professional Practice of Teachers*." Retrieved from http://www.doe.virginia.gov/administrators/superintendents_memos/2011/136-11.shtml

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

Darling-Hammond, L., & Bransford, J. (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do.* San Francisco, CA: Jossey-Bass.

Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.

Danielson, C. (2002). *Enhancing student achievement: A framework for school improvement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Donaldson, M. L. (2010). No more valentines. *Educational Leadership, 67*(8), 54-58.

Downey, D. B., von Hippel, P. T., & Hughes, M. (2008). Are "failing" schools really failing? Removing the influence of nonschool factors from measures of school quality. *Sociology of Education, 81*, 242-270.

Entwistle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools, and inequality.* Boulder, CO: Westview Press.

Ferguson, R. F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal of Legislation, 28*(2), 465-498.

Glass, G. V., Cahen, L. S., Smith, M. L., & Filby, N. N. (1982). *School class size: Research and policy*. Beverly Hills, CA: SAGE.

Glickman, C. D., Gordon, S. P., & Ross-Gordon, J. M. (2010). *Supervision and instructional leadership: A developmental approach* (8th ed.). Boston, MA: Allyn & Bacon.

Goldhaber, D. (2007). Everybody's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources, 42*, 765-794.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *Journal of Political Economy, 100*(1), 84-117.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Henry, G. T., Bastian, K. C., & Fortner, C. K. (2011). Stayers and leavers: Early-career teacher effectiveness and attrition. *Educational Researcher, 40*(6), 271-280.

Kane, T., Rockhoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 7-32). New York, NY: Teachers College Press.

Ladd, H. F. (2009). *Teachers' perceptions of their working conditions: How predictive of policy-relevant outcomes?* Working Paper 33. Washington, DC: CALDER, The Urban Institute. Retrieved from http://www.urban.org/uploadedpdf/1001440-Teachers-perceptions.pdf

Marzano, R. (2013). *Dr. Marzano's Teacher Evaluation Model aligned with his new School Leadership Evaluation Model*. Retrieved from http://www.marzanoevaluation.com/

McCutcheon, G. (1981). On the interpretation of classroom observations. *Educational Researcher, 10* (5), 5-10.

Measures of Effective Teaching. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Retrieved from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Measures of Effective Teaching. (2010a). *Classroom observations and the MET Project*. Retrieved from http://www.metproject.org/downloads/Classroom_Observation_092110.pdf

Measures of Effective Teaching. (2010b). *Content knowledge for teaching and the MET Project*. Retrieved from http://www.metproject.org/downloads/Teacher_Knowledge_092110.pdf

Measures of Effective Teaching. (2010c). *Student assessments and the MET Project.* Retrieved from http://www.metproject.org/downloads/Student_Assessments_92110.pdf

Measures of Effective Teaching. (2010d). *Student perceptions and the MET Project.* Retrieved from http://www.metproject.org/downloads/Student_Perceptions_092110.pdf

Measures of Effective Teaching. (2010e). *Teachers' perceptions and the MET Project*. Retrieved from http://www.metproject.org/downloads/Teacher_Perceptions_092110.pdf

Monk, D. H., & Haller, E. J. (1993). Predictors of high school academic course offerings: The role of school size. *American Educational Research Journal, 30*(1), 3-21.

Mosteller, F. (1995). The Tennessee study of class size in the early school grades. *The Future of Children, 5* (2), 113-127.

National Commission on Teaching and America's Future. (2003). *No dream denied: A pledge to America's children.* Washington, DC: Author.

New Teacher Project. (2007). *Hiring, assignment, and transfer in Chicago Public Schools*. New York, NY: Author.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, Rec. 1425. 115 Stat. (2002).

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26, 237-257.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin.

Peterson, K. D. (2004). Research on school teacher evaluation. *NASSP Bulletin, 88*(639), 60-79.

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

Pianta, R. C., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System*. Baltimore, MD: Paul H. Brooks.

Reardon, R. M. (2011). *Percentile growth modeling: A policy response to educational accountability*. Richmond, VA: Metropolitan Educational Research Consortium.

Rowan, B., 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. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement. Research Progress Report*. Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center.

Sawchuk, S. (2013). Multiple gauges best for teachers. *Education Week, 32*(17), pp. 1, 16.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Scriven, M. S. (1967). The methodology of evaluation. In *Perspectives of curriculum evaluation*, (American Educational Research Association *Monograph Series on Teacher Evaluation*, No. 1). Chicago, IL: Rand McNally.

Scriven, M. (1988). Duty-based teacher evaluation. *Journal of Personnel Evaluation in Education, 1*(4), 319-334.

Scriven, M. (1994). Duties of the teacher. *Journal of Personnel Evaluation in Education, 8*(2), 151-184.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher evaluation: Guide to effective practices.* Boston, MA: Kluwer Academic.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4-14.

Staiger, D., & Rockoff, J. (2010). Searching for effective teachers with imperfect information. *Journal of Economic Perspectives, 24*, 97-118.

Stronge, J. H. (2006). *Evaluating teaching: A guide to current thinking and best practice* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Stronge, J. H., & Ostrander, L. (1997). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating teaching: A guide to current thinking and best practice* (pp. 129-161). Thousand Oakes, CA: Corwin Press.

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

Virginia Department of Education. (2011). *Guidelines for uniform performance standards and evaluation criteria for teachers.* Richmond, VA: Author. Retrieved from http://www.doe.virginia.gov/teaching/performance_evaluation/guidelines_ups_eval_criteria_teachers.pdf

Virginia Department of Education. (2011). *Virginia standards for the professional practice of teachers.* Richmond, VA: Author. Retrieved from http://www.doe.virginia.gov/teaching/regulations/uniform_performance_stds_2011.pdf

Virginia Department of Education. (2011). *The research base for the uniform performance standards for teachers*. Richmond, VA: Author. Retrieved from http://www.doe.virginia.gov/teaching/performance_evaluation/research_base_ups_teachers.pdf

von Frank, V. (2011). Measurement makeover. *Journal of Staff Development, 32*(6), 32-39.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect*. Santa Cruz, CA: The New Teacher Project.