Learning from Teacher Observations:

Challenges and Opportunities Posed by New Teacher Evaluation Systems

Heather C. Hill and Pam Grossman

Harvard Graduate School of Education, Stanford Graduate School of Education

Author Note

The authors gratefully acknowledge funding from the National Center for Teacher Effectiveness (IES R305C090023), the National Science Foundation (1221693), and the Institute for Education Science (IES R305A110864).

Abstract

In this article, Heather Hill and Pam Grossman discuss the current focus on using teacher observation instruments as part of new teacher evaluation systems being considered and implemented by states and districts. They argue that if these teacher observation instruments are to achieve the goal of supporting teachers in improving instructional practice, they must be subject-specific, involve content experts in the process of observation, and provide information that is both accurate and useful for teachers. They discuss the instruments themselves, raters and system design, and timing of and feedback from the observations. They conclude by outlining the challenges that policy makers face in designing observation systems that will work to improve instructional practice at scale.

Learning from Teacher Observations:

Challenges ad Opportunities Posed by New Teacher Evaluation Systems

Teacher evaluation systems are undergoing a remarkable transformation. Spurred by strong federal incentives, most states have adopted procedures that combine data from student tests and rigorous observation protocols into scores intended for use in teacher accountability systems (Klein, 2012). Policy makers hope the new observation systems will replace those in which cursory evaluations have yielded little variation in observed teacher quality, and states and districts are now selecting observational instruments, training raters, and piloting these new systems in an effort to accomplish this goal. However, how states and districts will actually use scores from this new system is still unknown.

One use of scores that seems likely is as a source of information intended to help teachers improve instruction. Although authorizing legislation from almost all states includes potential consequences for poorly performing teachers, many policy makers have identified the direct, individualized teacher feedback from observations as the most promising feature in this new reform. A recent series of interviews with state policy makers found that more than half viewed teacher professional development and support as either the main or a major intended outcome of new teacher evaluation systems (Herlihy et al., in press), a view reflected in other recent commentaries and policy documents (Kane, McCaffrey, Miller, & Staiger, 2013; Papay, 2012). Thus, while observations were initially conceived as tools for evaluation, such protocols are now seen as key levers for the improvement of teaching. Hopes for using evaluation systems for improvement have been buoyed by promising results from early studies of observational feedback and coaching, which showed that both can positively affect student outcomes (Allen, Pianta, Gregory, Mikami, & Lun, 2011; Taylor & Tyler, 2011).

This optimism regarding the promise of evaluation for improving instruction is tempered by the history of education reform, where instructional improvement efforts often work at a small number of sites yet fail at scale (Elmore, 1996). Policy failure seems especially likely when the reform requires more than regulatory changes, attempts to change the core work of teaching, and is layered atop existing routines and practices (Cohen, 1988; Elmore, 1996; Murphy, 1989; Spillane & Zeuli, 1999; Tyack & Cuban, 1995). In the present case, all three of these conditions are met; teacher evaluation systems are designed to foster ambitious instructional change but will be grafted onto existing practices, including the use of generic observation instruments, the deployment of principals and other noncontent experts as primary observers, and the collection of only a small number of observations per teacher per year. We argue that these practices have proven unsuitable for generating and sustaining instructional improvement in the past and will fail to do so now if implemented as currently planned.

Instead, we suggest that to fulfill the potential of leveraging observation systems to support teachers in improving their practice, policy makers must build a complementary system for instructional improvement rather than assume that evaluation systems built for accountability can serve dual purposes. Based on our research and experiences observing past instructional improvement efforts, we argue that such a system for improvement must draw from research on both teaching and teacher development and must contain at least three features. First, these systems must make available subject-specific observation instruments that provide concrete guidance on desirable teaching practices. Second, these new systems must draw content experts within districts into the process of teacher evaluation, both for the sake of improving coherence in the messages transmitted to teachers and in order to leverage existing expertise around the improvement of instruction. Finally, states and districts must design systems in which feedback

from observations is both accurate and usable in the service of improving instruction. As these recommendations imply, such an observation system would prioritize teacher learning over accountability and also require numerous elements not contained in existing policy blueprints. For practical reasons, we imagine this goal would be best accomplished by building systems for improvement that complement existing systems for teacher evaluation.

New Meets Old

Historically, teacher evaluation systems based on principal observation have been ineffective at differentiating teaching quality. A survey of twelve districts and more than fifteen thousand teachers by Weisberg and colleagues (2009) described numerous problems, including that such systems in twelve districts failed to distinguish among teachers, rating most as proficient or highly proficient; that these ratings were typically based on two or fewer principal observations, each lasting well under an hour; and that districts seldom used scores from these observations in personnel decisions, including the granting of tenure and determining layoff priorities. Critically for current policy efforts, three-quarters of teachers surveyed reported that their most recent evaluations failed to identify areas for improvement, and almost half of teachers who did report an identified area for improvement stated that they failed to receive support for that work. Although no other sources provide a large-scale description of teacher evaluation, anecdotal reports suggest that this situation exists widely and has persisted for decades. To the extent that this is the status quo in many locations, reformers will need to think carefully and strategically about designing policies that support teacher learning. In particular, we recommend examining policy designs in three focal areas: instruments, raters, and the use of feedback.

Instruments

Content. Over the past two decades, most research into teachers, teaching, and teacher learning has taken place within discipline-specific communities. Although scholars have recognized the importance of general teaching skills, such as motivating student effort, maintaining a positive classroom climate, and maximizing instructional time, the majority of grant-funded and published research on teaching studied particular content areas—science, English language arts (ELA), and mathematics, to name a few. Recent policy has followed suit; the widely adopted Common Core State Standards identify discipline-specific competencies teachers must foster in students. Yet despite this prevailing view of teaching, most of the observation protocols selected in new teacher evaluation systems are generic with respect to content area and are designed to be used with all teachers—from kindergarten through calculus. Using generic instruments extends current evaluation practices. And from the perspective of policy makers, this makes sense: in a situation where every teacher must be evaluated on a regular basis, developing separate observation instruments for different subjects would be both time consuming and cumbersome.

In the service of finding expedient and efficient evaluation systems, however, we risk overlooking the importance of subject matter and the developmental needs of learners as they relate to teaching. The current systems ask us to believe that teaching kindergarten requires the same set of practices and knowledge needed to teach high school algebra. While all teachers may share common professional terrain—developing classroom routines to maximize learning time, representing content to a range of learners, establishing productive relationships with students—how they actually navigate these tasks depends, in large part, on the specific content they are teaching (Lampert, 2001; Shulman, 1987; Stodolsky, 1988). The expertise required of early childhood educators to establish routines for children just entering school—teaching beginning

literacy and numeracy and attending to the developmental needs of five year olds, among other competencies—differs from that required of the high school math teacher who must use fifty-minute blocks to help 150 or more students master the intricacies of algebraic thinking. In order to provide useful information for teacher learning, observational instruments will need to reflect these differences.

A specific example may help illuminate this issue. Questioning, and in particular using higher-order questions in instruction, is a practice included on many popular observation instruments. Yet understanding how best to tailor questions to the content is key. For instance, urging a teacher to "ask higher-order questions" may be generative in the case of some content but less fruitful in other areas. In mathematics, a lesson on simplifying fractions might allow the teacher to ask students to justify their answers or explain key mathematical ideas like equivalence. But a lesson on the order of operations, which is purely a mathematical convention, would not be as useful a site for such questions. Generic observation instruments would not allow such a subtle distinction, nor would they recognize that teaching a skill like the order of operations may allow teachers to make other mathematical points, such as the value of precision within the discipline. Similarly, assessing the ability of an ELA teacher to foster high-quality conversations around text or to support student writing through conferencing requires a complex representation of questioning unlikely to be captured by most observation protocols. Developing students' ability to think critically about text may require what Dennie Palmer Wolf (1987, p. 3) describes as an "arc of questions," which moves from literal questions about plot and character to more interpretive questions about what we learn about characters from what they say and do. Most generic instruments offer only blanket credit for higher-order questions, absent these finer distinctions.

Other practices that are specific to a particular subject matter, such as engaging students in high-quality investigations in science or using multiple primary-source documents to craft a historical argument, may be entirely missing from generic protocols, since these practices are not easily generalizable. Yet the ability of a science teacher to set up high-quality investigations that are tied to specific scientific phenomena is certainly an indicator of accomplished science teaching (Kloser, 2013), much as teacher skill in supporting the development of historical thinking and reasoning from conflicting documents suggests expert history teaching (Fogo, 2013; Wineburg, 2001). The absence of these practices from most observation instruments will limit the snapshot of teaching that emerges, the nature of feedback teachers are able to receive, and the diagnostic information districts can glean about subject-specific needs for professional development.

Grain size. One of the challenges for any observation instrument is getting the grain size right. By grain size, we mean the scope and level of detail around desired practices. While some observational instruments ask about practices that are fairly broad in their grain size (e.g., holding high expectations), others work at a smaller grain size, providing much more specification around such practices (e.g., pressing students for more complete and complex work; explicitly noting that students can achieve through effort). Grain size matters in both the design and use of observation instruments. The more specific the grain size, the more specific the feedback for teachers can be. However, in order to create an instrument that can work across multiple content areas and contexts, we suspect that many designers have been pressed toward more global descriptions of practice. One reason is that, as suggested above, evaluation systems designed for use across multiple grade levels and content areas must describe tasks in sufficiently broad strokes so that an instrument can work across settings. Another reason may be that it is

easier to get raters to agree on more general descriptions of practice; the more specification involved in a rating, the more the raters must understand the fine distinctions within a particular practice. However, efforts to improve instruction will almost necessarily entail decomposing these broader practices into more specific practices that teachers can tackle (Grossman et al., 2009).

Facilitating classroom discussions, for example, involves coordinating many different elements, including setting norms for discussion, selecting rich tasks or texts, asking appropriate questions, responding to student ideas, and getting students to respond to each other—all the while tracking the evolution of ideas throughout the discussion. For purposes of large-scale evaluation, it would be impossible to specify and evaluate each of these components; but in working with teachers to improve their classroom discussions, it is equally impossible to avoid digging into the details. Helping teachers increase the number of higher-order questions in mathematics classrooms may not only mean helping teachers to add more "why" questions to their repertoire but also involve supporting teachers to thoughtfully navigate student responses by using skills such as assessing the correctness and completeness of student comments, asking logical follow-up questions, and weaving student answers toward the larger mathematical point. These practices are too specific to place on any generic instrument, yet they are central to making real improvements in teaching.

This leads us to suspect that instrument grain size may represent an area in which using the same observation instrument for both teacher evaluation and feedback and improvement may prove difficult. Few policy makers, and certainly no teacher, would want an observational instrument that contains dozens of specific skills and behaviors, reminiscent of the teacher competency checklists that were popular in the 1980s. Keeping the grain size broad also may

enable raters to reach agreement more easily. However, using such instruments for teacher learning will require greater specification of the practices and attention to a range of practices that are subsumed into the broader indicators. Building a complementary system might thus entail creating tools that target more specific components of practice for improvement.

Raters and Observation System Design

Rater expertise. If we are right about the importance of content in the evaluation of teaching, then raters will require content expertise to use instruments correctly and to assist teachers in their improvement efforts. Clearly, content expertise is not always necessary: some elements of teaching—managing behavior, building a safe climate for learning, motivating student effort—are common across subjects, and, to the extent that a teacher requires assistance with these areas, a well-trained generalist observer might make a large difference in teaching and learning outcomes. However, as the above examples about content specificity imply, many elements in teaching do require either an understanding of the content or an understanding of how learners typically encounter the content.

Current policies would need to change in order to enable raters with content expertise to engage in teacher observations. A recent survey of state legislation and policy makers suggests that, as of now, most teachers will be evaluated exclusively or almost exclusively by their principals, similar to the teacher evaluation system that has existed historically (Herlihy et al., in press). Yet relying on the current system to help produce ambitious changes in instruction is risky. Principals are assumed by many to be generalists, and, in fact, evidence from studies of principals' views of mathematics instruction suggests that many lack the knowledge and expertise to provide content-specific feedback. For instance, in a study comparing secondary school personnel's teaching-related mathematical knowledge, principals and assistant principals

scored markedly lower than math teachers and coaches (Nelson, Stimpson, & Jordan, 2007). In another study, Goldsmith and Reed (in preparation) simulated principal feedback on instruction by asking 430 principals to comment on a scenario in which a teacher encouraged student debate over whether 5 can be divided by 39; in written comments, 40 percent of principals did not remark on the mathematics in the scenario, and another 25 percent made only cursory reference to the topic. Principals' level of attention to the mathematics was predicted by their mathematical knowledge, again suggesting that rater content expertise plays an important role in determining the feedback teachers receive. However, few evaluation systems require raters to have subject-specific expertise.

Accuracy and alignment of scores and feedback. Accuracy of scores is important for both accountability and improvement purposes. Yet, while accurate classroom-level scores are possible, state and district observation systems are not now designed to support high levels of accuracy (Bell et al., 2012; Hill, Charalambous, & Kraft, 2012; Kane et al., 2013). For instance, studies find that reliability rises when multiple raters view instruction from a specific teacher, yet in most states principals are designated as the sole observer (Herlihy, et al., in press). Evidence also suggests that reasonable reliability (where reliability is defined as systems that produce consistent scores regardless of raters and lessons sampled) can be achieved when raters view teachers on at least three or four occasions, yet this is more than the minimum legislated in many states.

The accuracy of scores matters both to evaluation and improvement efforts. For both ethical and legal reasons, personnel decisions must be based on accurate ratings of practice. However, inaccurate scoring also compromises the diagnostic function of observations. If score accuracy in the new systems is low, teachers may not receive important information about the

quality of their teaching and may miss opportunities to improve. To the extent that feedback stems from those scores, inaccurate scores may focus improvement efforts in the wrong area. For example, if observers inaccurately score teachers low on classroom management but do not detect problems in the accuracy of the content, a principal could end up providing unnecessary feedback on classroom management while ignoring the problems in representations of the subject matter.

Another area of concern around the utility of observational scores concerns their alignment with scores from value-added models, or teacher scores that are calculated based on student test outcomes. In theory, observational and value-added outcomes should align, yet in practice the evidence is mixed. While small-scale, within-year studies have generated evidence that observational and value-added outcomes correlate moderately (Bell et al., 2012; Grossman, Loeb, Cohen, & Wyckoff in press; Hill, Kapitula, & Umland, 2011; Schacter & Thum, 2004), the few large-scale studies that exist are less promising. In the Measures of Effective Teaching study, for instance, observers' instructional scores predicted prior teacher value-added scores at low levels (Kane & Staiger, 2012). If observational and student-test-based scores diverge in new systems within states and districts, then teachers might receive conflicting messages about improvement; the feedback on their own instruction may not relate to what they are incented to do based on student test scores. For example, teachers might receive high scores from value-added models but low scores on observational measures of the more ambitious instruction imagined by the Common Core State Standards. This constitutes a major problem that policy makers may grapple with over the next few years. If the new tests for the Common Core do indeed prove to be better aligned with state visions of teaching and learning, such a problem would be substantially ameliorated. Currently, however, the relationship between observation

instruments and state tests differs depending on the particular test used, and some assessments may do a better job of capturing intellectually challenging teaching and student outcomes, such as writing performance, better than others (Grossman, et al., in press). When observation instruments and test score data are not well aligned, policy makers must assist teachers by adopting assessments that reduce conflicting messages and bring expectations for teaching and learning into agreement.

Scores and Feedback Use

Time requirement. Currently, many state laws require between two and four observations per year for more experienced teachers, an incremental improvement over many old systems that required only one or two. In states already implementing new teacher accountability systems, such as Tennessee, many principals report that the higher number of observations is not feasible (Winerip, 2011). Yet opportunities for feedback, we suspect, must be even more highly concentrated to have a real effect on teaching.

Our hypothesis appears to be supported by evidence from research on coaching, which has been a popular intervention in teaching in recent years. Research suggests that coaching programs that are successful in supporting improved student outcomes provide at least monthly coaching sessions. In a recent study of the impact of literacy coaching, Biancarosa and Bryk (2011) report teachers engaged in one coaching session per month; in another study involving the Classroom Assessment Scoring System (CLASS) protocol for professional development, teachers completed between six and twelve coaching sessions (Allen et al., 2011). Ramey and colleagues (in press) assigned teachers to either weekly or monthly coaching. The study of the Cincinnati teacher evaluation system provides an exception to this rule; Taylor and Tyler (2011) found effects on student outcomes with a system that required four observations, mostly by

trained peers. While it is possible, based on these findings, that improvements in teaching can be seen with relatively small investments of time, most research suggests that the improvement of teaching is steady work requiring more than a few visits per year.

Feedback for improvement. Central to the idea of using observational data for improvement is the notion that feedback based on observation protocols can inform practice. In this sense, observation instruments become high stakes not just for personnel decisions but also for efforts to improve instruction. However, several issues related to feedback use may prevent data from observations from being integrated into plans for instructional improvement.

As noted above, the kind of feedback teachers receive will be framed largely by the categories of the observation instruments. And since most instruments are generic, this suggests that the feedback teachers are most likely to receive will also skirt the subject matter of instruction. Yet, research generally indicates that subject-specific professional development and coaching are more effective in improving instruction (Biancarosa, Bryk, & Dexter 2010; Cohen & Hill, 2001; Desimone, Porter, Garet, Yoon, & Birman, 2002). For example, improving student writing requires that teachers provide opportunities for students to develop a deeper understanding of the demands of a particular genre. The emphasis on argumentative writing in the Common Core ELA standards will require instruction that provides students with examples of what makes a good argument and with strategies for arguing with evidence and addressing counterarguments. Yet it's easy to see how feedback from an instrument that does not track the teaching of writing could sidestep such issues.

Further, the research on coaching suggests that individualized teacher feedback has been successful in cases where teachers are given specific, actionable items they can implement during their work with students (Biancarosa et al., 2010; Ramey et al., in press). As discussed

with respect to grain size, the language of most observational protocols does not describe such specific, actionable elements of practice. Many instruments contain lists of broad teaching characteristics rather than more specific (and, we assume, actionable) elements. Even observational instruments written at a higher level of specificity and which support relatively concrete teaching practices require translation between the language of the instrument and what teachers might do in classrooms. For example, the Protocol for Language Arts Teaching Observation (PLATO), designed specifically for English language arts, has elements that focus on the quality of modeling provided by the teacher. This would include a focus on the kind of metacognitive modeling teachers use to make visible strategies for reading or writing as well as the use of models of writing to illustrate specific features of a genre. But even modeling is composed of more specific teacher moves and practices, such as selecting pedagogically appropriate models, highlighting specific features of a model, using evidence to rebut a counterargument, or presenting some aspect of the writing process itself. To improve the teaching of writing, teachers need feedback that addresses their skill in these areas.

Finally, in the system as it is currently designed there is little room for follow-up and support after feedback is given. Feedback works best, we would imagine, when both parties are given a chance to check in and discuss progress made and problems encountered. Yet principals may have trouble scheduling follow-up conversations for purposes of feedback; they may not return to a teacher's classroom for months, at which time both may have forgotten important details surrounding the original feedback. Furthermore, we doubt that many districts have in place formal learning opportunities that move beyond the principal-teacher relationship itself to help teachers develop and practice the specific skills identified during feedback. Tailoring professional development to specific needs of teachers identified by observations is a promising

and often-mentioned area of reform; making this happen, however, will require redesigning professional development from the ground up as well as fostering closer alignment between instructional evaluation and support systems such as curriculum materials, observation protocols, and teacher learning opportunities.

Overcoming Challenges

We applaud the move to use new evaluation systems for the improvement of teaching. Moving the needle on student achievement will require districts to tackle the quality of instruction across teachers and schools, especially given the advent of higher standards for student learning. Observation instruments offer real potential for providing better feedback to teachers and policy makers on aspects of instruction that are ripe for improvement. However, realizing the potential of these instruments will require serious attention to the issues raised in this brief article, among others. If policy makers can begin planning now for ways to leverage evaluation systems for improvement, there may be a real opportunity to make a difference for teachers and students.

If the maxim that “we can only improve at scale what we can measure” holds true, then those invested in improving the teaching of content should be concerned about what’s missing from the observation protocols. In the desire to create one-size-fits-all systems of teacher evaluation, states and districts may risk relegating the substance of instruction to the sidelines. One way to address the limitations of more general protocols is to make subject-specific observational instruments available for use in certain situations. Teachers who are identified as struggling in a content area may benefit from observations that target the particular content area to help provide more specific feedback. Alternatively, if a district has prioritized implementing a reform such as the Common Core State Standards, it may choose to supplement generic

observation instruments with ones that reflect and support the discipline-specific nature of many of these teaching practices. Such work may be accomplished by finding instruments from the research sector, watching and waiting for Common Core–aligned instruments to be produced in the next several years, or collaborating with professional organizations and other districts to begin developing those instruments.

The challenges we detail here also suggest concentrating and differentiating resources to improve teaching. The relatively large number of observations needed to achieve ratings with acceptable reliability suggests that it may be more effective to focus scarce resources on teachers who are most in need of feedback. Conducting multiple observations on teachers who are struggling may yield more benefits than conducting fewer observations across all teachers. To accomplish this, policies might suggest different intervals of observations for teachers who have demonstrated their skill through previous evaluations; observing these teachers every two to three years may be sufficient. This would allow more time and resources to be devoted to high-quality observation and feedback for teachers who may most need targeted assistance.

Tackling the quality and content expertise of raters is another lever for maximizing the value of observations for both evaluation and improvement. In implementing any evaluation system, districts might explore how to identify raters with subject matter expertise from within existing resources in districts and schools. Some schools already have math and literacy coaches who bring expertise in these subject areas and who could be trained as observers for those subjects. At the secondary level, rethinking the role of department chairs to include observing and providing feedback could strengthen this role and provide opportunities for accomplished teachers to take on this responsibility. If there are concerns about department chairs observing their colleagues, districts could devise ways of having department chairs and coaches observe in

schools other than their own. Drawing district personnel with subject matter expertise into the new teacher observation system would simultaneously improve the coherence of messages reaching teachers—aligning observation-based feedback with existing curriculum and instructional priorities—and help ensure that at least some raters have a nuanced understanding of instruction in the content areas.

To leverage observation protocols more powerfully for instructional improvement, districts could also plan to produce what one might call "feedback bundles." Based on past observations and current district priorities, district staff may identify a small number of skills and practices critically needed in classrooms and then use observations to first assess whether specific teachers already possess those skills and, if not, to deliver a feedback bundle during the lesson debrief. Such a bundle would provide a rich description of the desired skill or practice, perhaps through a case study or video clip, but it would more importantly provide practice-based resources for getting the work done: information on how to find the practice or skill within the district's curriculum materials, names of teachers who might serve as resources during implementation, and connections between the skills and practices and state assessments. Such bundling of resources would make it more likely that new observation systems would lead to improvement and would support the work of principals, coaches, and professional developers within schools and districts.

As previous efforts to reform teaching remind us, policies that aim to support student learning must also create systems for teacher learning (Cohen & Hill, 2001). If districts are serious about leveraging observations for improvement, they will need to think longitudinally about how to continue to press for improvement around instructional practice. Changing practice is slow, steady work; in building systems for improvement, policy makers must resist the urge to

think that simply holding teachers accountable through evaluation systems will result in the changes in teaching that are required for students to meet more ambitious standards. Instead, policy makers must engage in the kind of high-demand, high-support policies that both help teachers learn more about the kinds of instruction envisioned by new standards and to receive the feedback and professional development required to develop new knowledge and skills.

If these new evaluation systems are to have a chance of improving the quality of teaching, policy makers must resist the urge to simplify the inherently complex nature of teaching. This will require grappling with how best to focus on issues of content in efforts to evaluate and improve teaching, how to select and develop raters with expertise in both observation and professional development, and how to concentrate resources to provide the kind of high-quality coaching that can actually impact practice. Such policies would be a serious departure from both existing and newly planned evaluation arrangements, but they would make it much more likely that teachers would receive the support and feedback they need to improve their own teaching and student learning.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*(6045), 1034–1037.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2–3), 62–87.

Biancarosa, G., & Bryk, A. S. (2011). Effect of literacy collaborative professional development. *Journal of Reading Recovery, 10(2)*, 25–32.

Biancarosa, G., Bryk, A. S., & Dexter, E. R. (2010). Assessing the value-added effects of literacy collaborative professional development on student learning. *Elementary School Journal, 111*(1), 7–34.

Cohen, D. K. (1988). *Teaching practice: Plus ça change*. East Lansing, MI: National Center for Research on Teacher Education, Michigan State University.

Cohen, D. K., & Hill, H. C. (2001). *Learning policy: When state education reform works.* New Haven, CT: Yale University Press.

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*(2), 81–112.

Elmore, R. F. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66*(1), 1–27.

Fogo, B. (2013, April). *Core practices for teaching history: The results of a Delphi panel survey.* Paper presented at the annual meeting of the American Educational Research

Association, San Francisco.

Goldsmith, L., & Reed, K. H. (in preparation). *Final report: Thinking about mathematics instruction NSF grant EHR 0335384*. Waltham, MA: Education Development Center.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record, 111*(9), 2055–2100.

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (in press). Measure for measure. *American Journal of Education.*

Herlihy, C., Karger, E., Pollard, C., Hill, H. C., Kraft, M. A., Williams, M., & Howard, S. (in press). State and local efforts to investigate the validity and reliability of scores from teacher evaluation systems. *Teachers College Record*.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Observational systems and a case for the G-study. *Educational Researcher, 41*(2), 56–64.

Hill, H. C., Kapitula, L. R., & Umland, K. L (2011). A validity argument approach to evaluating value-added scores. *American Educational Research Journal, 48*(3), 794–831.

Kane, T. J, McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Seattle: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching.* Seattle: Bill & Melinda Gates Foundation. Retrieved from http://www.metproject.org

Klein, A. (2012, June 16). More than half of states now have NCLB waivers. *Education Week.* Retrieved

from http://www.edweek.org/ew/articles/2012/07/18/36waivers.h31.html?tkn=ZSMFrarixRk6BCpX%2B9msZwJt%2FKiXlyOAiay0&cmp=clp-edweek

Kloser, M. (2013, April). *A different Common Core: An expert Delphi study on core science teaching practices.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lampert, M. (2001). *Teaching problems and the problems of teaching*. New Haven, CT: Yale University Press.

Murphy, J. (1989). Educational reform in the 1980s: Explaining some surprising success. *Educational Evaluation and Policy Analysis*, *11*(3), 209–221.

Nelson, B. S., Stimpson, V. C., & Jordan, W. J. (2007). *Leadership content knowledge for mathematics of staff engaged in key school leadership.* Newton, MA: Education Development Center.

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

Ramey, S. L., Ramey, C. T., Crowell, N. A., Grace, C., & Timraz, N. (in press). The dosage of professional development for early childhood professionals: How the amount, density, and duration of professional development may influence its effectiveness. In J. A. Sutterby (Ed.), *Early childhood professional development: Research and practice through the early childhood educator professional development grant*. Boston: The Emerald Group.

Schacter, J., & Thum, Y. M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*(4), 411–430.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23.

Spillane, J. P., & Zeuli, J. S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1–27.

Stodolsky, S. (1988). The subject matters: Classroom activity in math and social studies. Chicago: University of Chicago Press.

Taylor, E. S., & Tyler, J. H. (2011). *The effect of evaluation on performance: Evidence from longitudinal student achievement data of mid-career teachers* **(**NBER Working Paper No. 16877). Washington DC: National Bureau of Economic Research.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of American school reform*. Cambridge, MA: Harvard University Press.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project.

Wineburg, S. (2001). *Historical thinking and other unnatural acts: Charting the future of teaching the past*. Philadelphia: Temple University Press.

Winerip, M. (2011, November 6). In Tennessee, following the rules for evaluations off a cliff. *The New York Times.* Retrieved from http://www.nytimes.com/2011/11/07/education/tennessees-rules-on-teacher-evaluations-bring-frustration.html?pagewanted=all&_r=1&

Wolf, D. P. (1987). The art of questioning. *Academic Connections, 56*(3), 1-7.