Principal's Research Review

Supporting the Principal's Data-Informed Decisions

ISSN 1558-5948 VOL. 9, ISSUE 5 SEPTEMBER 2014

Teacher Evaluation Reform: Policy Lessons for School Principals

By Morgaen L. Donaldson and John P. Papay[1]

Over the past five years, teacher evaluation reform has swept the United States. Almost every state in the nation has revised its teacher evaluation policies, leading to substantial changes in how teacher evaluation is designed and implemented in schools and districts. In many ways, principals occupy the key position in these new evaluation reforms and their actions play a significant role in their success. What should school leaders know in order to manage the demands of new evaluation systems and leverage them to improve teaching and learning in their schools? This brief summarizes what is known about the measures used in current teacher evaluation initiatives and offers concrete steps that principals can take to facilitate their success.

Like all performance appraisal systems, teacher evaluation can serve at least two purposes. First, an evaluation system can provide accountability through monitoring, ensuring that teachers exert sufficient effort and follow specific standards and practices. It can also provide accountability by serving as a basis for rewarding or sanctioning teachers based on their performance. With better information about teacher performance, administrators can identify persistently low-performing teachers and dismiss them or counsel them to resign, and provide monetary or other rewards to high-performing teachers. Second, a strong teacher evaluation system can support teacher instructional development, offering actionable feedback to teachers about their practice and supporting them in improving their skills. Most evaluation systems embody these dual purposes. In this article, we examine how three widely used measures in today's evaluation systems support the dual purposes of teacher evaluation: standards-based observations, value-added measures, and student learning objectives.

Principals occupy the key position in these new evaluation reforms and their actions play a significant role in their success.

Standards-Based Classroom Observations

As of 2014, all states require that classroom observations be part of teacher evaluation ratings, making it the most commonly used component of teacher evaluation systems (Steinberg & Donaldson, in preparation). Unlike in the past, observation today typically involves assessing a teacher's instructional practices against an articulated set of performance standards. Formal observations are generally structured around an observation instrument or protocol that helps evaluators measure teachers' performance by translating observed practice on specific standards into detailed descriptors within a rubric.

Most states and districts that have developed professional standards for teachers typically adapt their standards and protocols from existing models. Charlotte Danielson's Framework for Teaching is the most popular model, but several other examples exist. Of

nassp national association of secondary school principals

these, the National Board for Professional Teaching Standards (NBPTS) are also used regularly, while some states and districts have sought to incorporate work from more specialized standards and observational instruments for English language arts (Protocol for Language Arts Teaching Observation), mathematics (Mathematics Quality of Instruction), and classroom interactions (the Classroom Assessment Scoring System).

Existing research on these standards-based protocols documents widespread variation in teacher practice and shows that teachers who score higher on these rubrics also have greater estimated contributions to student test scores (i.e., value-added). Although test scores and observations may measure different dimensions of practice, examining their relationship proves useful. Standards-based observation ratings actually predict student test performance above and beyond that teacher's value-added rating (Kane et al., 2013).

As tools for developing a clear measure of teacher productivity for accountability purposes, standards-based evaluations face several tradeoffs. First, for better or for worse, they rely on human judgments of practice. This allows supervisors to develop a broad-based understanding of a teacher's practice, thereby allowing them to make potentially stronger inferences about a teacher's true contributions to student learning than measures based on test scores permit. However, standards-based evaluations are only as good as the assessments that evaluators make. Evaluators report that it is difficult to separate what they know of the teacher, or the teacher's contributions outside of the classroom, from their judgments of the teacher's instructional practice (Donaldson, 2013; Papay & Johnson, 2012). Administrators consistently assign higher evaluation ratings to teachers than do external observers or than they do on informal effectiveness ratings provided to researchers (Harris et al., 2014). Although having clear standards, using highly qualified and well-trained evaluators, and focusing on evidence can help remove much of the subjective bias in observation measures, separating the personal from the professional can be difficult.

Standards-based evaluations are only as good as the assessments that evaluators make.

A second consideration is that standards-based observations take substantial human resources to complete. Time constraints often require evaluators to make judgments based on a relatively limited sample of instruction, which can limit their reliability. Evaluators are typically school principals, who have many other responsibilities and often struggle to find sufficient time to do this work well. Although building and maintaining reliability takes a substantial investment, it is possible to achieve. Incorporating multiple observations into an evaluation helps a great deal, particularly if the observations are unannounced. Here, briefer walk-throughs might be warranted, particularly if coupled with more substantial observations at several points during the year. And, training for evaluators can be invaluable. For example, in Cincinnati, evaluators generally observe at least four complete lessons before making their final determinations, and all evaluators must complete a rigorous training program, earn certification as reliable observers, and participate in ongoing training and professional development for evaluators (Johnson et al., 2010). Research also suggests that using multiple observers can resolve some of these potential reliability challenges (Kane et al., 2013). To resolve the cost implications of this decision, some districts, like New Haven, CT, require multiple evaluators only for teachers with very high or very low ratings that could result in rewards or sanctions (Donaldson & Papay, 2014).

For the purpose of teacher development, standards-based observations have strong potential. They can provide direct and specific feedback about a teacher's instructional practice. Observation reports and summative evaluations include detailed information about teachers' strengths and weaknesses and thus can be meaningful to teachers and administrators. The performance standards lay out overall expectations for

performance and rubrics describe a continuum of instructional quality, thereby offering teachers a clearer sense of the practices they need to adopt in order to improve their practice. Outside of evaluation, rubrics and standards can also provide fodder for broader conversations between administrators and teachers and among teachers about how to improve instruction (Donaldson et al., 2014). In this way, evaluation systems with standards-based observations potentially enhance teachers' skills, thereby serving the evaluation's development purpose.

For the purpose of teacher development, standards-based observations have strong potential.

Value-Added Measures

Over the past decade, the federal government has pushed states and districts to include student achievement measures in teacher evaluation, and states have responded: 80% of states implementing new evaluation systems require that teachers be evaluated at least in part on student achievement (Steinberg & Donaldson, in preparation). Value-added measures, which use test scores to estimate a teacher's contribution to student achievement, are a popular but controversial option. In theory, supervisors want information about teacher productivity, and value-added methods can provide some evidence in this regard. However, some researchers view them as a cause for concern (e.g., Baker et al., 2010), primarily because they may not account well enough for differences in the types of students that teachers teach and may not be sufficiently reliable to estimate accurately a teacher's contributions to student learning. In large part, this debate depends on how value-added data will be used to hold teachers accountable.

The central idea behind the different types of value-added models involves comparing the test scores of a teacher's students to the predicted scores had the students had an "average" teacher. Models that use student growth percentiles are quite similar to value-added models as we discuss them here. Analysts make a variety of decisions in estimating these predicted levels of achievement, such as whether to account for student characteristics and whether to compare teachers in the same school or across the district. These are, in effect, policy decisions without clear answers. For example, if the models account for the disadvantages of student poverty, they essentially hold the teachers of low-income students to different standards than those of higher-income students. However, if they do not account for poverty, they may not fully control for out-of-school factors affecting students.

In the past decade, substantial research has examined value-added models. By and large, research finds that, on average, students whose teachers had higher value-added scores have better long-term outcomes (Chetty, Friedman, & Rockoff, 2013). Importantly, though, value-added estimates only capture one dimension of a teacher's performance: Jackson (2014) finds that measures based on test scores alone understate the total effect of teachers on students.

While these findings suggest that value-added measures, on average, can identify meaningful performance differences among teachers, estimates for individual teachers can vary widely, for example, across the specific statistical model or test used, or over time. Two central issues emerge here. First, do value-added models provide an unbiased estimate of the student achievement gains caused by a teacher? Second, are value-added estimates reliable enough to provide sufficiently precise estimates of teacher effectiveness?

In terms of bias, value-added models are based on standardized tests, which are typically graded by machine or by external scorers with no chance for systematic favoritism towards certain teachers. But, simply removing such favoritism does not mean that value-added models are free of bias. Bias can arise if the students assigned to some teachers are systematically different from those assigned to others. For instance, if a school principal assigns several disruptive children to a teacher she believes can best serve these students, that teacher's value-added might not

National Association of Secondary School Principals

G.A. Buie
President

Michael Allison
President-Elect

JoAnn Bartoletti
Executive Director

Jennifer J. Jones
Director of Communications

Lisa Schnabel
Graphic Designer

Principal's Research Review is a publication of NASSP, 1904 Association Dr., Reston, VA 20191-1537. Telephone 703-860-0200. Fax 703-476-5432. Website www.nassp.org.

NASSP dues include the annual subscription rate of $50; individual subscriptions are not available. NASSP members can download this issue at www.nassp.org/prr.

reflect a fair comparison with other teachers in the school. Researchers have examined whether value-added models can fully account for differences in student assignments across individual teachers, but this debate has not been resolved (Kane et al., 2013; Chetty et al., 2013; Rothstein, 2010).

A broader (and less studied) question is whether value-added models can isolate an individual teacher's causal impact on student performance (Reardon & Raudenbush, 2009). For instance, attributing a student's mathematics test gains to her mathematics teacher may not fully recognize the contribution that a science teacher made to the student's mathematical knowledge. And, it is often difficult to determine which students should be attributed to which teacher, because of within-year student mobility, formal or informal team teaching, and data system limitations.

The second main concern involves the precision of these estimates. Value-added estimates necessarily include some margin of error around a teacher's score, and in many cases, this error is sufficiently large that only the very top and bottom performers can be differentiated statistically from the average teacher. As a result, researchers, even those who advocate the use of value-added approaches, often call for using multiple years' worth of data to construct an estimate for an individual teacher (Glazerman et al., 2010; Baker et al., 2010).

As a measure to inform teacher development, value-added measures also have substantial limitations. Teachers typically receive a summative score, with no information about their performance on specific areas of instructional practice and no targeted feedback about how to improve. Thus, value-added, as currently conceived, is not well aligned with evaluation as a tool for professional growth.

In addition, several practical limitations also restrict the broad-based application of value-added models regardless of the evaluation purpose. Of greatest concern, given standardized testing practices in most states, only a minority of teachers work in a grade or subject area that supports value-added analysis (typically English language arts and mathematics teachers in grades 4–8). Second, these estimates are only as good as the tests on which they are based. Because many state tests are designed to measure whether students are proficient on a set of standards, they may not be particularly good for evaluating teachers. Student performance on tests does not reflect the full range of learning that educators and parents care about. Finally, because value-added estimates rely on the test data, principals need to wait at least until test results are available during the summer before assessing performance. And, any efforts to include multiple years of data to have sufficiently reliable estimates further limit their use for annual teacher evaluation.

Student Learning Objectives (SLOs)

Some districts and states seeking an alternative to value-added measures for incorporating student achievement data in teacher evaluation have adopted student learning objectives (SLOs). SLOs are goals set by teachers and administrators for each classroom based on existing assessments. Teachers are then evaluated based on how well their students meet these goals. The key, of course, is in how goals are set. Importantly, the rigor of a teacher's goals depends both on the individual teacher and on his/her interactions with the principal, not some external standard.

The use of SLOs to evaluate teachers is relatively new. Denver, CO, became one of the first districts to use SLOs in 1999 when it adopted them as a key component of ProComp, its performance-based pay system for teachers. SLOs have been adopted as evaluation measures in 24 states (Steinberg & Donaldson, in preparation). However, little research has explored their use in teacher evaluation (Goe & Holdheide, 2011). Two studies of Denver's four-year pilot program found that teachers who met their goals had somewhat higher student growth scores than teachers who did not (Goldhaber & Walch, 2012; Slotnik et al., 2004). However, over time, teachers became better able to meet the goals they had set, with 98% of experienced teachers meeting both their objectives during the final year.

Nonetheless, SLOs have several strengths. First, unlike value-added measures, SLOs can apply to all teachers and all specialists. Second, goal-setting theory suggests that individuals are motivated by goals that they set for themselves (Locke & Latham, 2002). This finding is most relevant to SLOs' use as a development tool, because they may encourage teachers to build skills and improve their instruction. Teachers in New Haven reported that SLOs were the most positive aspect of their new evaluation system and helped them focus their instruction on student performance (Donaldson, 2013). Furthermore, teachers report that the adoption of a new evaluation system incorporating SLOs led them to spend more time analyzing student data and assessing student progress than they had prior to the new system (Donaldson et al., 2014). Third, these measures, which are jointly constructed by teachers and administrators, may increase teachers' buy-in and investment in their own evaluation (Briggs, 2013).

Other Means of Evaluating Teachers

Although standards-based evaluations and test-score measures have received the most attention, other means of evaluating teachers are gaining some currency in systems across the country. Research on these measures and their potential to support teacher accountability or development is limited. Student surveys are being used in an increasing number of districts (Hull, 2013). Research has found that students are able to distinguish a teacher whom they like from one who promotes their learning and that student perceptions are correlated with teacher performance (Kane et al., 2013). States such as New York, Connecticut, and Utah and districts such as Miami-Dade, FL, are also using parent surveys/feedback in teacher evaluation. The little research that exists suggests that parent survey data may contribute to improving measures of teacher performance (Peterson et al., 2003).

Parent survey data may contribute to improving measures of teacher performance.

Some districts include portfolios, consisting of teachers' instructional materials and student work, in teacher evaluation. Secretary of Education Arne Duncan has publicly recognized Tennessee's Fine Arts Growth Measures System as a model for how states can evaluate teachers of non-tested subjects like art (Robelen, 2013).

Outcomes of Current Teacher Evaluation Systems

Systems that incorporate standards-based observations, value-added measures, student learning objectives, and other types of measures are relatively new, and the effects of these evaluation systems have not been widely studied. Nonetheless, a growing body of high-quality research suggests that rigorous evaluation systems have had important effects.

First, evaluation systems can spread out the distribution of ratings of teacher performance and lead to more teachers being dismissed or counseled out of the profession. Historically, a very small percentage of teachers have received low evaluation ratings; the great majority of teachers have received "satisfactory" or better ratings. Even fewer teachers have been dismissed due to poor performance. Evidence suggests that new evaluation systems have increased

these percentages. For example, in 2011–2012, after evaluation reform in Washington, DC, 16% of the city's teachers received the highest rating and 15% received a rating in one of the lowest two categories, while departure rates of teachers with low ratings were substantially higher than in earlier years (Dee & Wyckoff, 2013).

More importantly, participating in a rigorous evaluation system can improve teacher practice and raise student test scores, helping to build teacher skills in a lasting way. Taylor and Tyler (2012) found that participating in Cincinnati's standards-based evaluation system improved the performance of teachers not only during the year they were evaluated, but also in subsequent years. Research from Chicago and Washington, DC, also suggests that participating in evaluation systems can improve teacher effectiveness (Steinberg & Sartain, forthcoming; Dee & Wyckoff, 2013). Together, these three studies suggest that teacher evaluation can lead to substantial and lasting improvements in teacher performance, both in districts that attach incentives to teacher evaluation ratings (e.g., DC) and those that do not (e.g., Cincinnati).

Three studies suggest that teacher evaluation can lead to substantial and lasting improvements in teacher performance.

Conclusion

Since 2008, changes to teacher evaluation have occurred on an unprecedented scope and scale. The challenge for principals comes in understanding what the evaluation measures can—and cannot—tell us about teacher effectiveness, and coupling accountability with support and opportunities for teachers to learn. Teacher evaluation is a prime policy lever as a conduit to combine accountability and support; leaders can use the information embedded in these systems to provide valuable feedback to teachers and help structure individualized learning opportunities for them.

Participating in a robust evaluation system, even one without particularly high stakes for teachers, can improve an individual teacher's effectiveness over time (e.g., Taylor & Tyler, 2012), suggesting that the teacher evaluation development pathway is likely a promising approach to improving teacher performance at scale. States and districts recognize this opportunity in rhetoric, but, to date, few have made progress in designing new evaluation systems to promote instructional improvement. Instead, this burden often falls on principals, who are asked to implement these new requirements and make them work well for all teachers in their schools. This reality leads to four main implications for principals as they tackle this work.

First, taking on the dual challenge of evaluating teachers and providing actionable feedback for improvement is understandably daunting for many principals. To be effective, evaluators need to be able to provide feedback and design processes to help teachers use this feedback to improve their instruction. Doing all of this work effectively requires expertise that few principals have. One promising strategy to address this challenge is to leverage teacher expertise in providing feedback. Cincinnati, OH, and Montgomery County, MD, have been using expert teachers as peer evaluators for more than a decade, with notable success (Johnson et al., 2010; Papay & Johnson, 2012). Both districts have robust teacher evaluation systems that, as noted above, have produced impressive results. More broadly, for evaluation to achieve its goals, state and district leaders need to prioritize support for principals, and principals need to ask for the support they need.

Second, principals can push district and state leaders to improve the implementation of these evaluation systems to reflect on-the-ground realities. By and large, policymakers are only beginning to tackle the challenges of how to use evaluation to support teacher development (Papay, 2012). Here, principals can inform this work at the state and district level. For example, efforts in some states to develop peer

networks among principals to share best practices seem promising. Local involvement can help policymakers recognize the challenges involved in this work and the support necessary to do it well.

Third, principals need to work to understand the measures on which evaluation is based. All measures have important strengths and limitations. Knowing what measures can say about a teacher's practice, and what they cannot, is critical to determining how best to use evaluation data to inform instructional improvement both for individual teachers and for the faculty schoolwide.

Finally, principals can invest in the evaluation process. As educators are well aware, policy reform in education often happens in cycles, with reforms moving in and out of favor quickly (Hess, 1999). However, teacher evaluation is not merely a district or state policy, but also a school-based strategy necessary to improve teacher quality. Despite the inherent growing pains that will occur with any new system, a stronger evaluation system is a critical element for enhancing the skills of educators in a school. By investing in building a robust system of teacher evaluation and development, school leaders can bolster the culture of professional learning in their school and, with time, students will reap the benefits of improved instruction.

[1] This article draws from a more extensive review of teacher evaluation research and reforms in Donaldson, M.L. & Papay, J. (forthcoming). Teacher evaluation for accountability and development. In (Eds.) H.F. Ladd & M.Goertz, Handbook of Research in Education Finance and Policy.

References

- Baker, E.L., Barton, P.E., Darling-Hammond, L., Haertel, E., Ladd, H.F., Linn, R.L., & Shepard, L.A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper 278. *Economic Policy Institute*.
- Briggs, D.C. (2013). Teacher evaluation as Trojan horse: the case for teacher-developed assessments. *Measurement: Interdisciplinary Research and Perspectives*,11(1-2),24-29.
- Center on Great Teachers and Leaders. (2014). National Picture: A Different View. Retrieved March 31, 2014 from http://www.gtlcenter.org/sites/default/files/42states.pdf.
- Chetty, R., Friedman, J.N., & Rockoff, J.E. (2013). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. National Bureau of Economic Research Working Paper 19423.
- Dee, T., & Wyckoff, J. (2013). Incentives, selection, and teacher performance: Evidence from IMPACT. National Bureau of Economic Research Working Paper 19529.
- Donaldson, M.L. (2013, April). "How Do Teachers Respond to Being Evaluated Based on Their Students' Achievement? Evidence from New Haven, CT." Paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.
- Donaldson, M.L., Cobb, C., LeChasseur, K., Gabriel, R., Gonzales, R., Woulfin, S., & Makuch, A. (2014). *An Evaluation of the Pilot Implementation of Connecticut's System for Educator Evaluation and Development*. Storrs, CT: Center for Education Policy Analysis.
- Donaldson, M.L. & Papay, J.P. (2014). An Idea Whose Time Had Come: Negotiating Teacher Evaluation Reform in New Haven, Connecticut. Paper presented at the annual conference of the American Educational Research Association, Philadelphia, PA.
- Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). Evaluating teachers: The important role of value-added. Brown Center on Education Policy, Brookings Institution.
- Goe, L., & Holdheide, L. (2011). Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects. Washington, DC: NCCTQ.

■ Goldhaber, D. & Walch, J. (2012). Strategic pay reform: A student outcomes-based evaluation of Denver's ProComp teacher pay initiative. Economics of Education Review 31, 1067-1083.
■ Harris, D.N., Ingle, W.K., & Rutledge, S.A. (2014). How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal*, 51(1),73-112.
■ Hess, F.M. (1999). *Spinning Wheels: The politics of urban school reform*. Washington, DC: Brookings Institution Press.
■ Hull, J. (2013). Trends in Teacher Evaluation: How States are Measuring Teacher Performance. Alexandria, VA: Center for Public Education.
■ Jackson, C.K. (2012). Non-cognitive ability, test scores, and teacher quality: Evidence from 9th grade teachers in North Carolina. National Bureau of Economic Research Working Paper 18624.
■ Johnson, S.M., Papay, J.P., Fiarman, S.E., Munger, M.S., & Qazilbash, E.K. (2010). *Teacher to Teacher: Realizing the Potential of Peer Assistance and Review*. Washington, D.C.: Center for American Progress.
■ 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. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.
■ Locke, E.A., & Latham, G.P. (2002). Building a practically useful theory of goal setting and work motivation: A 35 year odyssey. *American Psychologist*, *57*, 705-717.
■ Papay, J.P. (2012). Refocusing the Debate: Assessing the Purposes and Tools of Teacher Evaluation. *Harvard Educational Review*, *82*(1), 123-141.
■ Papay, J.P., & Johnson, S.M. (2012). Is PAR a good investment? Understanding the costs and benefits of teacher peer assistance and review programs. *Educational Policy*, *26*(5), 696-729.
■ Peterson, K.D., Wahlquist, C., Brown, J., & Mukhopadhyay, S. (2003). Parent Surveys for Teacher Evaluation. *Journal of Personnel Evaluation In Education*, *17*(4), 317-330.
■ Robelen, E.W. (2013). Teacher-Review Tool: Classroom Portfolios. *Education Week*, *33*(4), 1-20.
■ Rothstein, J. (2010). Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *Quarterly Journal Of Economics*, *125*(1), 175-214.
■ Reardon, S.F., & Raudenbush, S.W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy*, *4*(4), 492–519.
■ Slotnick, W.J., & Smith, M.D. (2004). *Catalyst for Change: Pay for Performance in Denver Final Report*. Boston: Community Training and Assistance Center.
■ Steinberg, M. & Sartain, L. (forthcoming). Does teacher evaluation improve school performance? Experimental evidence from Chicago's Excellence in Teaching Project. *Education Finance and Policy*.
■ Taylor, E.S. & Tyler, J.H. (2012). The Effect of Evaluation on Teacher Performance. *American Economic Review*, *102*(7). 3628-3651.