Teacher Evaluation Models: Policy Research

Erin Borthwick, Sarah Cohodes, James Sennette, and Andrea Touhey

Assuming that the goal of a teacher evaluation system is to differentiate teachers in order to improve teaching practices, appropriately target professional development to them, counsel out ineffective teachers, and/or strategically place or promote effective teachers all in the service of improving student achievement, we can evaluate teacher evaluation systems in their ability to do the above. Specifically, in order to investigate teacher evaluation systems and how they affect teachers, schools, and students, we ask three questions about them:

1. How is the teacher evaluation system implemented?
2. Does the teacher evaluation system differentiate between teachers?
3. Does the teacher evaluation system influence teacher practices and/or student achievement?

We will investigate teacher evaluation generally through this framework, as well as some selected teacher evaluation programs.

Of course, it is possible that differentiating teachers is inappropriate or impossible, and that the vast majority of teachers perform at the same level. Almost any student or parent, however, can tell stories about a particularly influential and effective teacher as well as school years stuck in a classroom with a subpar teacher. This anecdotal evidence is also reflected in an extensive body of research on the effects of teachers on student test scores. In a well known example using evidence from Tennessee, where a longitudinal data system made possible the tracking of students over time, researchers have been connecting teachers to the *value added*

they contribute to student achievement scores (Sanders & Rivers, 1996). This methodology calculates an expected test score for students based on their prior test score performance and personal characteristics. If their actual test score is different from the expected test score, this difference is attributed to the teacher. If a teacher facilitates student learning so that most her students score higher than their expected test score, that teacher would have high value added. If a teacher's instruction leads to her students scoring lower than their expected test score, that teacher would have low value added. While this methodology has some critics that point to the unreliability of using high stakes standardized tests scores for teacher ratings (McCaffery, Koretz, & Lockwood, 2003), it has been experimentally validated as unbiased (Kane & Staiger, 2008). There is also preliminary evidence that these value-added measures are connected to specific teacher practices, indicating that growth in student test scores may be connected to differentiated teaching practices, rather than (or perhaps in addition to) some aspect of teacher personality (Tyler, et al., 2009). Thus, given that teachers do appear to have different effects on their students, it is reasonable to suggest that an evaluation system can differentiate teachers and identify which teachers should be targeted for professional development, dismissal, or other programs such as leadership development.

Teacher Evaluation Systems Nationally

Most states or school districts require some form of teacher evaluation, either through state law or specific teacher contracts. Nationally, school districts and states have implemented a variety of approaches; however, most of these entities use a common approach. Typically, teacher evaluation consists of principal observation of a teacher's practice with a rating of that teacher on a binary "satisfactory" or "unsatisfactory" scale. Some school districts have more

detailed scales that might include levels like “unsatisfactory,” “basic,” “proficient,” and “advanced.”

Evaluation is implemented differently across school districts, but the typical observation based evaluation system requires little time spent in the classroom by the evaluator (often the principal), no specific evaluation training for the evaluator, and a rubric that may focus on superficial aspects of teaching, such as employee dress and attendance, rather than instruction (Toch & Rothman, 2008). As such, Toch and Rothman refer to these types of observations as “drive-bys” and consider them not taken seriously either by the administrator evaluators or the teachers being evaluated. Three teachers and administrators from Texas, where the state system only requires a single 45 minute observation, all note that the evaluation process would be more successful if more time was spent on it and there were multiple observations (personal interviews with Brown, G., November 6, 2009, Miller, F., November 12, 2009, and Wallace, J., November 11, 2009). A recent report by The New Teacher Project, The Widget Effect, also notes that these cursory evaluations are coupled with a culture where all teachers expect to get high ratings (Weisberg, et al., 2009). The Center for American Progress calls this the “Lake Wobegon effect,” where all teachers are above-average (Donaldson, 2009). In such a culture, it is rare for evaluations to accurately reflect the quality of a teacher’s instruction.

While the typical observation based evaluation system might not be incredibly detailed or carefully implemented, it is possible that principals can correctly identify effective and ineffective teachers based on brief observations of their practice. This may be the case, however, it is not reflected in the ratings that principals give teachers: when using a satisfactory/unsatisfactory scale, less than one percent of teachers are rated unsatisfactory; when using a more detailed scale, still less than one percent of teachers are rated unsatisfactory and 94

percent of teachers are rated in one of the top two categories (Weisberg, et al., 2009). With such a small number of teachers rated unsatisfactory, current teacher evaluation systems do not appear to differentiate teachers in order to target programs or policies to them. In addition to the under-identification of unsuccessful teachers, the practice of rating all teachers towards the top of the scale means that truly extraordinary teachers are also not identified and thus leadership roles, bonuses, and incentive programs cannot be targeted to the very top of the teaching profession.

Given this, it is not surprising that research does not connect these types of systems to changes in teacher practice or increased student achievement. Donaldson (2009) identifies seven reasons why typical evaluation systems have not improved teaching and learning: 1) the "Lake Wobegon effect" makes it difficult to fire poor teachers; 2) even if evaluation systems and instruments identify poor teachers, other forces prevent principals from pushing from using those results to fire teachers; 3) many evaluation instruments (like rubrics and forms that principals use to evaluate teachers) do not make it easy to differentiate teachers and may not be aligned with district or school instructional focuses; 4) district policies, like additional paperwork around low ratings, stop principals from rating teachers poorly; 5) training on evaluation tools is poor; 6) evaluation processes do not typically focus on feedback to teachers, and thus they do not know what or how to change their practices; and 7) "evaluation has few consequences, positive or negative." And while the ratings on some teacher evaluation instruments that are more sophisticated than an satisfactory/unsatisfactory binary correlate with gains in student test scores (Wilson, et al., in press; Milanowski, Hallman, & White, 2004; Tyler, et al., 2009; and Pianta, 2005),[1] the evaluation systems or ratings have not been connected to *increased* student

[1] Wilson, et al. (in press) connect the Connecticut BEST evaluation system ratings to higher reading gains. Milanowski, Hallman, & White (2004) find rubrics based on a Danielson framework also connect to test score gains in Cincinnati and Las Vegas and Tyler, et al. (2009)

achievement. The best evaluation systems may accurately identify teachers, but the implementation of the evaluation system itself does not appear to have changed student achievement trajectories, most likely for the reasons listed above. Although it would be difficult to isolate the effect of a teacher evaluation system on student achievement, since evaluation systems may be implemented as part of other school reforms and are typically implemented district wide, it would be possible to test if different evaluation systems randomly assigned to different schools in a district increased or did not change student achievement.

Findings on Selected Teacher Evaluation Systems

Cincinnati Teacher Evaluation System

The Cincinnati Teacher Evaluation System (TES) was developed jointly by the Cincinnati Public School district and local teacher's union in the late 1990's. It uses an evaluation rubric based on the Danielson framework. Cincinnati evaluators are exemplar teachers from the system that have received special training from the district and leave the classroom for three years to evaluate and coach their peers. Evaluators are matched by subject and grade level and observe teachers multiple times per year. TES differs from the typical binary evaluation systems with its fine grained rubric, intense training for evaluators, reliance on multiple observations, and professional development focus and support.

Cincinnati's TES is also one of the few teacher evaluation systems that has been connected to student achievement growth (Milanowski, Hallman, & White, 2004). Recent preliminary research goes into more detail about the connection between TES and student

---

confirm the Cincinnati finding. Pianta (2005) connects a rubric he developed to test score gains as well.

achievement growth: Tyler et al. (2009) generally find that high TES scores correspond with high value-added and more specifically, they suggest, based on their findings, that "given teachers who have similar proficiency in 'teaching practices'…, helping teachers improve their 'classroom environment' management … will likely also generate higher student achievement (the TES rubric is available at: http://www.cps-k12.org/employment/tchreval/stndsrubrics.pdf). Third, given two teachers who are equally adept at 'content and standards focused teaching,' the teacher who adds 'inquiry-based pedagogy' practices will generate higher reading achievement, but not higher math achievement." Thus not only do they correlate TES scores with student achievement gains, but they are able to focus on specific teacher practices that are associated with those gains.

However, neither Cincinnati study investigated TES's impact on student achievement as a whole in the district, likely due to the difficulty of separating the effect of TES on student achievement from other policies and programs in the late 1990's and 2000's. As such, there is no indication that TES improved student achievement in Cincinnati, perhaps due to the difficulty in measuring its impact, perhaps due to the dilution of its impact through the mechanisms described in the Center for American Progress report (detailed above). Even this seemingly exemplary system rated only 6.5 percent of teachers unsatisfactory in the teaching for learning domain (Weisberg, et al. 2009). A researcher from the recent study that associated TES ratings with student achievement, E. Taylor, also noted that expecting to see large changes in student achievement by simply measuring teacher practice might be unrealistic due to the difficulty in changing the practices of the teachers being evaluated. He noted that "reading the Cincinnati rubric makes it seem like it would be easy to move from "proficient" to "distinguished" -- there's only one point on the rating scale between these two categories, and the difference in evaluative

language seems tractable -- but the standard deviation of the TES score is half a scale point. That means empirically to move from 'proficient' to 'distinguished,' [a teacher has] to move two standard deviations, which is hard" (personal interview, November 9, 2009). While an evaluation system is able to reflect a teacher's current instructional practice, it doesn't communicate the complexity of advancing to the next level.

Teacher Advancement Program System

The Teacher Advancement Program (TAP) is a multi-faceted system for managing and improving teachers in schools. One element of the TAP system is a proprietary rubric that is used, in addition to growth scores based on student achievement, to evaluate teachers. These evaluations are related to TAP's professional development program, but also connect to its performance-based compensation. TAP teachers are evaluated multiple times per year by trained evaluators. The evaluation element of TAP has not been studied separately from the entire TAP program; thus it is difficult to attribute effects of TAP to the evaluation system or the program in general. However, teachers that participate in the TAP program have been shown to have higher student growth than teachers in comparison schools that did not use the TAP program (Solmon, Cohen, & Woo, 2007).

Texas Professional Development and Appraisal System and *Boston Public Schools Performance Evaluation of Teachers System*

Unlike the Cincinnati or TAP system, neither the Texas Professional Development and Appraisal System (PDAS) nor the Boston Public School's (BPS) teacher evaluation system has been extensively studied. As such, we turn to our interviews with Texas and Boston educators to determine how the systems are implemented. In Texas, F. Miller, an assistant principal,

described evaluations limited by time and by fear on the part of those being evaluated (personal interview, November 12, 2009). One BPS teacher described the evaluation process at multiple schools and concluded "[evaluation] depended on who the evaluator was" with some evaluators being cursory and others having more extensive and helpful processes (personal interview, November 12, 2009). Without specific research on the PDAS and BPS systems, it is impossible to determine if the evaluation tools successfully differentiate between effective and ineffective teachers and if the evaluation process improves student achievement. PDAS uses a rubric where teachers are rated "Exceeds," "Proficient," "Below," or "Unsatisfactory" and BPS uses a rubric where teachers are rated "Satisfactory" or "Unsatisfactory." Given their use of typical rubrics, it is likely that both the PDAS and BPS systems encounter the many problems described above that teacher evaluations encounter throughout the nation – few teachers identified as unsatisfactory, little differentiation, cursory evaluation sessions, and little response to or effects from the evaluation.

References

Donaldson, M.L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality*. Center for American Progress. Available: http://www.americanprogress.org/issues/2009/06/teacher_evaluation.html

Kane, T., & Staiger, D. (2008). Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation. NBER Working Paper No. 14607. *National Bureau of Economic Research.*

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

Milanowski, A. Kimball, S.M., & White, B. (2004). The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites. CPRE-WU Working Paper. *Center for Education Research, University of Wisconsin-Madison.*

Piata, R. (2005). Spotlight: Classroom observation, professional development, and teacher quality. *The Evaluation Exchange*. 6(4). Available: http://www.gse.harvard.edu/hfrp/eval/issue32/spotlight3.html.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee Value-Added Research and Assessment Center.

Solmon, L.C., White, T.J., Cohen, D., & Woo, D. (2007). The effectiveness of the Teaching Advancement Program. *National Institute for Excellence in Teaching.* Available: http://www.fldoe.org/dpe/pdf/effectiveness-of-TAP.pdf

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

Toch, T. & Rothman, R. (2008, January). Rush to judgment: Teacher evaluation in public education. Education Sector. Available: http://www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2009). Using student performance data to identify effective classroom practices. Brown University and Harvard Graduate School of Education. Working Paper. Available: http://www.hks.harvard.edu/pepg/PDF/events/colloquia/Tyler_Colloquia_Working_Paper.pdf

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New York: The New Teacher Project. Available: http://www.widgeteffect.org.

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