EPIC Education and the Public Interest Center
University of Colorado at Boulder

EPRU Education Policy Research Unit
Arizona State University

| | |
|---|---|
| DOCUMENT REVIEWED: | "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences" |
| AUTHORS: | Daniel Weisberg, Susan Sexton, Jennifer Mulhern, and David Keeling |
| PUBLISHER/THINK TANK: | The New Teacher Project |
| DOCUMENT RELEASE DATE: | June 1, 2009 |
| REVIEW DATE: | August 5, 2009 |
| REVIEWERS: | Raymond L. Pecheone and Ruth Chung Wei |
| E-MAIL ADDRESS: | pecheone@stanford.edu; rchung@stanford.edu |
| PHONE NUMBER: | 650-723-4106; 650-723-8399 |

SUGGESTED CITATION:
Pecheone, R. L. & Wei, R. C. (2009). *Review of "The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences."* Boulder and Tempe: Education and the Public Interest Center & Education Policy Research Unit. Retrieved [date] from http://epicpolicy.org/thinktank/review-Widget-Effect

## Summary of Review

*The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences*, published in June 2009 by the New Teacher Project, examines how 12 school districts across four states use teacher evaluation to make human resources decisions. It then proposes how to build teacher evaluation systems that are more credible and useful. Overall, the report portrays current practices in teacher evaluation as a broken system perpetuated by a culture that refuses to recognize and deal with incompetence and that fails to reward excellence. However, omissions in the report's description of its methodology (e.g., sampling strategy, survey response rates) and its sample lead to questions about the generalizability of the report's findings. In addition, while the rationale for the report's policy recommendations is sound, the proposals are restricted to the findings from the study and fail to consider or to draw upon any promising teacher evaluation strategies in current use. Transforming the system rather than tinkering around the edges will require broader thinking and a commitment to provide much greater investment and support for innovation to build, test, and audit evaluation systems that can stand up to public scrutiny and be practically feasible.

# Review

## I. INTRODUCTION

The new report from The New Teacher Project, *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences,*[1] begins with the well-supported premise that teacher quality is one of the strongest determinants of student achievement. If teacher quality and instructional effectiveness are the keys to improving student achievement, it is logical to examine how school systems use teacher evaluation to identify top talent, to strengthen teacher capacity and, when warranted, to dismiss teachers who are not competent.

Through the use of mixed methods, including surveys of key stakeholders (teachers and administrators), semi-structured interviews, and teacher evaluation performance data within and across 12 districts in four states, the report concludes that the evaluation practices currently in use are fundamentally flawed and indefensible. The extraordinary lack of variation across districts paints a portrait of teacher evaluation yielding inaccurate findings of equal effectiveness and universal competence, a problem perpetuated by a culture that treats teachers as essentially interchangeable parts—widgets.

If it is true that almost 100% of teachers who are rated satisfactorily or receive the highest rating possible are truly performing satisfactorily, the report asserts that we would expect to see more of the schools in the 12 selected districts to be meeting their Annual Yearly Progress targets. However, as the Widget report highlights, even in districts where 98% of teachers were rated satisfactorily (e.g., Denver Public Schools), many schools failed to meet their AYP targets (p. 12). Following this line of reasoning, there should be little differentiation in a teacher's impact on student learning (as represented by test scores) regardless of the teacher of record. However, decades of research on teacher quality (including recent studies cited by the report) clearly indicate that there are vast differences in teacher effectiveness. Some teachers produce relatively rapid growth in student learning while others produce good to fair gains, and for a significant proportion of teachers, there is little or no evidence of impact on student learning.

The Widget report exhaustively chronicles flaws in the ways in which teacher evaluation is conducted in the 12 districts studied. The report further asserts that while teacher evaluation systems espouse lofty goals to improve practice and hold high standards for the profession, the data suggest the opposite. The process of teacher evaluation is so entangled in negotiations around due process and minimal standards ("do no harm") that evaluation often is compliance-driven, perfunctory, and devalued to the point where districts invest little time, money and resources in training administrators, supporting teachers, or pushing poor teachers to get better or leave the profession. Moreover, current compensation structures pose a major barrier to attracting and retaining high-quality teachers because of compressed pay scales that are based on years of experience rather than effectiveness. Existing recruitment, retention, tenure, and dismissal policies are inimical to improving teacher quality. The report concludes that overcoming the "Widget Effect" is imperative to improve the quality of instruction and to ensure equitable outcomes for all students, and the authors propose a series of recommendations to arrive at evaluation systems that are

credible and defensible—systems that teachers, policymakers and parents can trust.

As discussed below, the report's findings are consistent with what has been found by other experts. However, the lack of information about the study's sampling strategy and response rates on surveys limits readers' ability to generalize the findings and also leads to questions about why the report chose not to sample the states and districts that are known for more rigorous teacher evaluation approaches (e.g., Florida, South Carolina) or for providing incentives for effective teaching (e.g., districts with career ladder programs, performance-based compensation[2]). In addition, the report fails to rest its policy recommendations on research that documents the features of effective, educative, valid, and reliable evaluation structures and strategies. The report appears to generally ignore the research base on teacher evaluation, and makes no mention of existing appraisal systems that hold high standards for examining teacher effectiveness. Close examination of these evaluation systems might provide important lessons and direction for how to reform the current system of evaluation. Finally, and as detailed below, the report's recommendations echo past rhetoric of teacher evaluation critiques and do not go far enough to question the current constraints of the system. Meaningful change will require systems thinking, and a commitment to provide much greater investment and support for innovation to build, test, and audit evaluation systems that can both stand up to public scrutiny and be practically feasible.

## II. FINDINGS AND CONCLUSIONS OF THE REPORT

The study profiles the health status of teacher evaluation practices and declares without reservation that the patient—teacher evaluation systems—is in critical condition and suffering from the "Widget Effect": an indifference to variations in teacher performance. As evidence of this, the report provides a set of findings (summarized on the report's page 6) that illustrate the problem:

- Evaluation ratings are uniformly high (superior or distinguished) across all demographic sub-groups, including years of experience (ranging from 93% in Chicago to 82% in Rockford) (p.11).
- Teacher self ratings of their performance was also uniformly high (84% of all teachers rated their performance at 8 or more out of 10 points) (p.22).
- Novice teachers appeared to receive no special attention or extra support (66% of novice teachers are rated superior or better and 76% are very confident that they will receive tenure after the probationary period has ended) (p.15).
- Most teacher evaluation systems are based on 2 or fewer observations totaling about 75 minutes or less (64% of tenured teachers and 59% of probationary teachers were observed 2 times or fewer) (p.21).
- Evaluators spend approximately the same time on feedback and coaching regardless of whether teachers were highly or poorly rated (56% of highly rated teachers and 58% of poorly rated teachers reported that they received feedback on their performance) (p.21).
- Survey results indicate that administrators received limited or no formal training in evaluation. The number of teachers dismissed or judged as needing improvement was about the same regardless of whether administrators had received training (p.22).

Based on the flaws in evaluation systems as presented here, we do not find it surprising that judgments of teacher quality are made

primarily on the basis of longevity and credentialing rather than on instructional quality.

## III. THE REPORT'S RATIONALE FOR ITS FINDINGS AND CONCLUSIONS

The report's primary findings are based on its data collection in the 12 districts studied, and the report provides four recommendations based, at least in part, on these findings. The first recommendation**—Adopt a comprehensive performance evaluation and development system that fairly, accurately and credibly differentiates teachers based on their effectiveness in promoting student achievement and provides targeted professional development to help them improve**—is essentially a response to the flaws identified in current teacher evaluation systems and documented throughout the report. While the recommendation seems to place a focus on the measurement of "effectiveness in promoting student achievement," the report does acknowledge limitations of value-added models for measuring student learning growth across the grade levels and subject areas (p. 27). The report therefore suggests that it would be possible to assess a teacher's effectiveness in promoting student achievement through an observation-based evaluation process, with value-added modeling as a useful supplement to the core evaluation process.

The report also asserts that evaluation systems should differentiate between more and less effective teaching so that schools can act on these differences in teaching performance, whether through dismissal, retention, targeted professional development, salary differentials, or other forms of recognition or career advancement. The underlying reasoning here and throughout the four recommendations is that the overall quality of teaching would be improved by changing the population of teachers (through dismissals as well as voluntary exit of incompetent teachers), by motivating teachers to perform more effectively through salary rewards or other forms of recognition, and by providing a means to identify teachers in need of additional support..

The first recommendation also includes several features that the report assumes would make performance evaluation systems more effective: clear and straightforward performance standards; multiple, distinct rating options; regular monitoring and norming of evaluators; frequent and regular feedback to teachers; professional development targeted to individual teacher needs; and intensive support for teachers falling below performance standards. These features are cited by the report as being absent from the evaluation systems that they studied. The first three are aimed at improving the technical quality of evaluation instruments and building the capacity of evaluators to make credible and reliable judgments so that the evaluation system can validly differentiate between more and less effective teaching. The last three features are focused on the subsequent uses of the evaluation to inform formative learning opportunities for teachers identified as needing additional support. The fundamental rationale for improving the accuracy of evaluation tools and processes is to provide the means to improve the quality of teaching by providing accurate feedback about areas for growth, to identify teachers in need of support, and to provide targeted professional development in specific areas of weakness.

The second recommendation—**Train administrators and other evaluators in the teacher performance evaluation system and hold them accountable for using it effectively**—is based on the report's finding that teacher evaluation thus far has been perfunctory (since almost 100% of teachers usually receive satisfactory ratings) and has not led to fair, reliable, or credible judg-

ments of teaching performance. The logic underlying this recommendation is that even with high-quality evaluation instruments and processes, many supervisors responsible for using these instruments will continue to inflate their ratings without adequate training on how to rate reliably against a rigorous set of standards. In addition, to counter the asserted culture of indifference about less-than-satisfactory performance, evaluators need to be held accountable for how they utilize the evaluation process to improve the quality of their teaching staff. The accountability piece is critical to this argument because even well-trained supervisors may not have the motivation to rate teachers using rigorous standards without some stake in the outcomes of these ratings.

The third recommendation—**Use performance evaluations to inform key decisions such as teacher assignment, professional development, compensation, retention and dismissal**—is based on the premise that evaluations, when fair, reliable, and credible, should be attached to some stakes for teachers so that teachers and evaluators take the evaluation process seriously and in order to "create cultures of excellence in schools, where the focus is on achieving individual, group and school performance goals related to student achievement" (p. 29). In addition, as mentioned previously, there is an underlying logic that the quality of teaching and student learning will be improved by excluding the least effective teachers, motivating satisfactory teachers to improve their effectiveness through rewards and recognition, and identifying and addressing weaknesses through targeted professional development.

The fourth and final recommendation—**Adopt dismissal policies that provide lower-stakes options for ineffective teachers to exit the district and a system of due process that is fair but streamlined and efficient**—is a response to the very low rates of teacher dismissal that the report documents across the 12 districts. The underlying logic seems to be that given systems in which dismissal is usually reserved only for teachers who endanger children and involves a lengthy, expensive legal process, very few teachers are actually dismissed, even when they receive unsatisfactory ratings. The perspective of the report is that dismissals based on a fair, reliable, and credible evaluation process will become routine and accepted, which will lead to a culture in which poorly rated teachers voluntarily "bow out" of the profession or their schools when policies facilitate exit and provide incentives for incompetent teachers to exit their positions.

## IV. THE REPORT'S USE OF RESEARCH LITERATURE

Two substantive areas of research literature are cited in the report: 1) research that supports the idea that the quality of teaching has a significant impact on student learning and achievement,[3] and 2) research on the usefulness and limitations of value-added approaches to measuring teaching effectiveness.[4] These research studies are used appropriately to support the report's rationale for its argument to strengthen teacher evaluation systems. However, the report cites neither the research literature on the issues and challenges surrounding teacher evaluation in general, nor research on specific approaches to teacher evaluation—approaches that have been successful in discriminating between more and less effective teachers, that are educative, and that provide incentives for instructional improvement (these are summarized briefly below).

The absence of these key pieces of research on teacher evaluation diminishes the report's arguments and recommendations because

they are consequently based solely on this new survey research. They therefore stand alone, divorced from any sense of other contexts around the challenges and affordances of various evaluation approaches, as well as the political, organizational, and economic ramifications of designing and implementing rigorous, fair, reliable, and credible teacher evaluation systems. For example, the proposed remedy of improving the training of evaluators to rate teaching more reliably and holding them accountable for how they use evaluation to improve teaching makes no differentiation between principals, department chairs, or peers, versus external evaluators, concerning their relative effectiveness as fair, objective, reliable raters. Nor does it consider the relative difference between using evaluators with or without subject- or grade-specific pedagogical expertise. The recommendation also does not take into account the level of time and budgetary investments required to attain appropriate levels of reliability in evaluator ratings or to provide the kind of regular monitoring that is proposed. In sum, the recommendations for what a rigorous and educative performance evaluation system should include are made without grounding them in additional research about approaches that would support teacher learning and ultimately make a positive difference in teaching and student learning.

Several genres of research on teacher evaluation might inform proposals for how to reinvent evaluation systems and that also raise questions for future research:

**Research on the *content* of teacher evaluation instruments.** The report offers a satisfactory critique of the lack of differentiation that current instruments make between more and less effective teaching. However, it completely ignores the content of those evaluation instruments and how they actually define effective teaching. Kennedy points out that teacher evaluation instruments

> …have not attended to the intellectual substance of teaching: to the content actually presented, how that content is represented, and whether or how students are engaged with it. This may seem like a surprising and glaring omission, especially since it has been pointed out more than once [citations omitted]. But it is not surprising at all when we consider the difficulty of agreeing on the meaning to the events we see. And any assessment of the intellectual and substantive merits of teaching is entirely about its meaning.[5]

Kennedy also notes that evaluation instruments relying on observations alone will always be insufficient because they do not allow the evaluator to get at the full meaning of teachers' actions in the classroom.

Another weakness of existing instruments is that we do not know whether most of them are actually valid indicators of teacher effectiveness or predictors of student learning. The research literature the authors could have turned to has examined this and could have helped provide information and perspective. Several research studies, for example, document the predictive validity of teacher ratings on instruments used for research (e.g., CLASS, IQA).[6] Additional research documents the correlation between a district-level teacher assessment and student achievement gains.[7] The Widget Effect report simply accepts the content or construct validity of the instruments used in any of the districts surveyed. But policymakers attempting to apply the report's recommendations should question and explore the validity of those instruments.

**Research on the impact of who implements teacher evaluation.** Past research has called into question the reliability of teacher evaluations conducted by principals and administrators. In particular, some emerging research explores this question of who is most qualified to evaluate the effectiveness of teachers and compares—for example, the ratings of principals, teachers, and students.[8] The ineffectiveness of current administrator-led teacher evaluation systems leads us to question whether this model can attain a sufficient level of reliability to be credible. From a formative point of view, some research suggests that beginning teachers rate more highly the support received from mentors with the same grade/subject matter backgrounds.[9] Moreover, higher new teacher retention has been associated with having mentors in the same subject area.[10] This research highlights the importance of the pedagogical expertise needed to make fine-grained and insightful judgments about a teacher's performance. This sort of evaluation is more likely to go beyond surface-level characteristics of teacher performance such as classroom management, engagement of students, organization of classroom resources, and time management.

**Research on the role of teacher evaluation in dismissals.** Pullin[11] reviews the research on teacher dismissals and surveys a number of legal cases in which teachers were dismissed. She presents the research of Walsh-Sarnecki,[12] who documents the perceptions principals and administrators have of the barriers to the dismissal process due to legal and monetary requirements. That research also explains the efforts of principals and administrators to avoid the painful process of filing dismissal proceedings, by transferring teachers or by offering incentives to leave. While it appears that courts are more likely to side with districts in such legal proceedings,[13] the number of teachers dismissed as a result of a poor performance evaluation has been negligible.[14] Pullin finds in her survey of dismissal cases that the causes for dismissal are generally for "clearly wrong and uncontroversial" infractions such as "drinking beer with cheerleaders, serious sexual misconduct, [and] changing student responses on a test…" rather than the quality of pedagogical practice.

One of the recommendations proposed by The Widget Effect is to use evaluation ratings as the basis for dismissal, or to trigger the provision of low-stakes options for teachers to remove themselves from their positions or to move to other schools. However, other recent empirical studies[15] suggest that low performers are already self-selecting themselves out of the profession. According to this research, the least effective teachers (as determined by value-added measures) were the most likely to transfer schools or to leave teaching altogether (without being formally dismissed). Thus, the low numbers of formal teacher dismissals cited by the report may reflect the fact that teachers who have received unsatisfactory evaluation ratings are already being given low-stakes options for leaving their positions voluntarily, which often go unrecorded as outcomes of low evaluation ratings.

**Research on the role of teacher evaluation in tenure decisions.** Sykes and Winchell (in press)[16] provide a framework for understanding the role of teacher evaluation in tenure decisions. They also critique the use of value-added measures to evaluate teacher effectiveness because of technical and practical problems with this methodology, as well as because they favor evaluation methods that build teacher professionalism and organizational capacity by involving teachers themselves. For such reasons, the use of value-added measures even as a supplement to teacher evaluations for making tenure de-

cisions is understood in the report to be problematic.

Further, Sykes and Winchell cite a recent survey study that reports the vast majority of teachers are dissatisfied with current observation-based assessments for tenure: "Almost seven in 10 teachers (69 percent) say that when they hear a teacher at their school has been awarded tenure, they think that it's 'just a formality—it has very little to do with whether a teacher is good or not.'"[17] Estimates from an earlier 1998 survey report published by the American Federation of Teachers (AFT) and the National Education Association (NEA)[18] also indicate that most teachers agree that about 5% of their colleagues are "poor teachers." This estimate of the percentage of poorly performing teachers is confirmed by other studies, with some indicating that up to 10% of teachers may perform less than satisfactorily.[19] (These figures are fairly consistent with what the Widget Effect found through their surveys of teachers and administrators, reported on page 18.) Sykes and Winchell also cite a study by Marshall,[20] who found that while teacher unions were once the champions of tenure, the "new unionism" has advocated for the profession to become more involved in teacher evaluation for tenure decisions through Peer Assistance and Review (PAR) processes. This suggests that in some contexts there has already been a shift in the culture of evaluation in teacher professional communities, and a growing sense of need for more accurate and credible evaluation for tenure review. Yet this "new unionism" and cultural shift goes unnoticed and unmentioned in the Widget report.

**Research on promising performance-based models of teacher evaluation.** The report does not include any mention or discussion of research on models of teacher evaluation that have shown promise for achieving greater levels of fairness, reliability, and credibility, as well as for promoting teacher learning and development. Marshall notes that a growing number of teacher union locals, including those that have joined the Teacher Union Reform Network (TURN), as well as these locals' districts, have followed Toledo's example of PAR models of teacher evaluation.[21]

Two statewide models of teacher induction and evaluation processes are also worth noting. California's Formative Assessment and Support System (CFASST) and the New Teacher Center's Formative Assessment System (FAS) are both part of the state's Beginning Teacher Support and Assessment program. In Connecticut, the Connecticut Competency Instrument (CCI) and the Beginning Educator Support and Training (BEST) portfolio both evolved from the Interstate New Teacher Assessment and Support Consortium (INTASC), a performance assessment project sponsored by the Council of Chief State School Officers. Both state models of teacher evaluation are based on observations of teacher performance by mentor teachers as well as the use of teaching artifacts to support the evaluation.

The research review by Youngs, Pogodzinski, and Low[22]—focused on the PAR teacher evaluation systems, the California Beginning Teacher Support and Assessment and the Connecticut Beginning Educator Support and Training induction programs—includes information on these models and their impact on teacher learning and performance.[23] Another promising model comes from the Milken Family Foundation's Teacher Advancement Program (TAP), a career advancement program that operates in more than 130 schools across 14 states and the District of Columbia. The TAP combines evidence from four evaluations annually by a master/mentor teacher and evi-

dence of student learning through value-added measures to evaluate teaching effectiveness. It then provides opportunities for professional learning based on teacher evaluation measures, as well as performance-based compensation. A 2007 evaluation study found that teachers in TAP schools outperformed comparison groups on value-added measures.[24, 25]

## V. Review of the Report's Methods

The report's methods include the following: (1) surveys of 1,300 administrators, 15,000 active teachers, and 790 former teachers across 12 districts in four states; (2) examination of teacher evaluation records; evaluation instruments; teacher dismissal, transfer, and attrition records; district and state policies regarding teacher evaluation; and collective bargaining agreements; and (3) 130 interviews of district leaders, school board members, human resources staff members, legal counsel, labor relations specialists, union leadership, school principals, other evaluators, and teachers. The report includes a careful analysis of the states' and districts' teacher evaluation policies and how teacher evaluation factors into human resources decisions (e.g., hiring, tenure, dismissal, and remediation). In addition, information on the districts' evaluation instruments and processes (such as frequency of observations) as well as teacher evaluation data (such as distribution of ratings, number and percentage of teachers receiving unsatisfactory ratings, and number and percentage of teachers dismissed.) are reported and are used by the authors to build a compelling case for the inadequacy of the current teacher evaluation instruments, processes, uses, and policies in the four states.

There are two critical omissions in the report, however. These concern the selection of the sample of states and districts and concerning the response rates on surveys, both of which affect readers' ability to infer the representativeness of the sample and the generalizability of the findings. The report states that the sample represents states that employ diverse teacher performance management policies and have demonstrated a commitment to improving teaching and learning. Districts were selected to represent diverse size, geographic location, evaluation policies and practices, and "overall approach to teacher performance management" (p. 5). Yet it is unclear how and why particular districts were selected, and whether the sample captures the range of teacher evaluation practices being implemented in school districts and states across the United States. In fact, it is curious that the report did not include states/districts with a reputation for having more rigorous teacher evaluation policies and practices (e.g., Florida and South Carolina), "right to work" states with weak teacher unions (e.g., Georgia, Texas, and most Southern states), states with strong incentives for National Board certification (e.g., North Carolina and South Carolina), or states that have implemented performance-pay (e.g., Louisiana and South Carolina) or career ladder programs (e.g., Arizona). Selecting states/districts that have contrasting policies on teacher evaluation would have enabled the researchers to observe how truly different approaches to evaluating teachers evaluation and to *using teacher evaluation results* play out in terms of the quality of the states' and districts' teacher evaluation approaches, instruments, processes, and uses for making human resources decisions. This, in turn, might have affected the report's recommendations.

## VI. Review of the Validity of the Findings And Conclusions

The findings of this report are not surprising and have been documented over the last several decades by other researchers.[26] Further, states have shown little leadership in support-

ing and promoting teacher quality, as evidenced by the fact that only 14 states have legislation that requires school systems to evaluate teachers at least once a year.[27] These so called "drive-by" evaluation practices often involve nothing more than superficial judgments about teacher behavior based on ratings of teacher competence as either acceptable or unsatisfactory, with no attention paid to whether their students are learning. It is not surprising that such evaluation practices are less than educative for teachers.

Even though this report's sample of districts seems to stack the odds against finding examples of more rigorous evaluation practices, the larger picture of teacher evaluation practices in the U.S. is not totally bleak. There are a number of teacher assessment initiatives that move beyond perfunctory evaluations and authentically focus on instruction and student learning. Some notable examples (cited above) include the Teacher Advancement Program (TAP) supported by the Milken Family Foundation, the INTASC/Connecticut Beginning Educator Support and Training Program (BEST) and the National Board for Professional Teaching Standards (NBPTS).

TAP is an intensive evaluation and support program patterned after Charlotte Danielson's model to improve teaching by focusing on instruction and student learning. Through an investment in coaching and creating positive incentives, the TAP evaluation system is specifically linked to programs that support career ladders and performance-based compensation. TAP is in operation in over 130 schools across 14 states and the District of Columbia.

The Connecticut BEST program is a statewide evaluation system focusing on beginning teachers (first three years of teaching). BEST was established in 1989 by statute to improve teacher quality by providing new teachers with mentors and training and requiring all teachers to submit a portfolio of their teaching, which is then evaluated by subject-area peers. All novice teachers in Connecticut must meet performance standards for teaching by year 3 to be eligible for a provisional teaching license. The BEST system is a multiple-measures system focusing on capturing teacher and student data around a unit of instruction, including evidence of planning, instruction (video tape), assessment and teacher thinking. In short, BEST attempts to capture a body of evidence through multiple measures on how teachers develop student understandings of a topic over time, including evidence of student learning. Moreover, a value-added study of the BEST model found that performance on BEST was significantly correlated with student scores on the state accountability test (students of high-performing teachers outperformed a matched sample by 4-6 months of growth in the area of literacy).[28]

The NBPTS certification process is also a multiple-measures system that includes lesson plans, instructional materials, student work, video of teachers working with students, teacher reflections, and evidence of work with parents and peers. Since 1987 the NBPTS has granted certification of accomplished teaching to more than 63,000 teachers in 16 subject areas. Some states, like North Carolina, have made heavy investments to encourage teachers to pursue certification, and a large proportion of teachers in the state have applied for and achieved certification. Additionally the NBPTS is an example of an evaluation system that provides important professional learning experiences. While the findings on the relationship between certification and student achievement are mixed, the body of evidence on National Board certification

documents significant improvements in candidates' instructional practices[29] and encouraging results with regard to the usefulness of certification as a signal of teacher quality.[30]

The models cited above highlight promising teacher assessment practices and illustrate the benefits of development of teacher quality while maintaining high standards. They also provide examples of subject-specific, multiple-measures evaluation instruments (scored by subject and grade-specific instructional experts) that rely on both classroom observation (e.g., through a videotape) as well as artifacts of teaching (e.g., lesson plans, assignments and student work samples) as the body of evidence for assessing the meaning and content of teachers' instruction. Finally these evaluation systems all involve peer review with scorers who are highly trained to make independent, reliable judgments about teacher quality. As mentioned earlier regarding the literature review, the report's conclusions and recommendations would have been strengthened through consideration of this knowledge base.

## VII. USEFULNESS OF THE REPORT FOR GUIDANCE OF POLICY AND PRACTICE

The picture the report paints of the landscape of current practices in teacher evaluation is an indictment of a broken system perpetuated by a culture that refuses to recognize and deal with incompetence. The report's careful march through the data collected by the authors leads to findings consistent with conclusions drawn by other reports about the current state of affairs in teacher evaluation.[31] The rationale for the report's policy recommendations is sound, and the recommendations appear to represent reasonable strategies for improving a broken teacher evaluation system.

However, these proposals echo the past rhetoric of teacher evaluation critiques that have not resulted in systemic reform of teacher evaluation, and the proposals do not move beyond the constraints of the current paradigm of administrator-led teacher evaluation that relies on observation alone. Multiple measure systems of assessment supported by subject-specific peer review seem to be more rigorous, more research-supported, and more valid.[32]

Another policy perspective that may strengthen the report's recommendations focuses on systems thinking around reform policies that affect districts and states. Teacher evaluation is only one small part of a comprehensive strategic system designed to manage human capital. Human capital in this instance refers to how an organization acquires, supports, and retains top talent over time, addressing the full continuum of policies and practices that impact teachers over their entire careers. As the private sector has learned, the highest performing organizations not only recruit and maintain top talent but also manage them in ways that support the strategic direction of the organization, including human resources functions such as recruitment, screening, selection, equitable placement, induction, professional development, evaluation, and compensation and promotion into instructional leadership roles.[33]

The expressed goal of this sort of systems thinking is to reinvent the management of human capital around the dual goals of building trustworthy metrics of student learning and teacher performance. Al Shanker, in a 1985 speech entitled "The Making of a Profession," called for a conceptual framework that included differentiated teacher pay, peer review, and a national framework for the assessment of teaching and learning.

Building upon such systems thinking approaches to human capital as well as learning from the promising evaluation models cited in this review leads to a few additional policy recommendations:

1. Develop and support a joint state and federal initiative to create voluntary core standards for teaching that are aligned to the national core student standards. Without a common definition of effective teaching that is aligned with learning objectives for students, the content of instruments used to assess teaching effectiveness will continue to vary in those instruments' ability to differentiate more or less effective instructional practices linked to student learning and achievement.
2. Provide state and federal resources to support the creation of innovation zones for the development of new models for teacher evaluation and the strategic management of human capital. Schools within these innovation zones, much like charter schools, could be exempt from aspects of collective bargaining agreements that constrain hiring, evaluation, tenure, and retention policies.
3. Develop technology platforms to support the on-line training of raters and the scoring, calibrating, benchmarking, and reporting of teacher performance data through independent peer and administrator review.
4. Establish a district or state board of examiners that will review and approve evaluation programs and practices. The Board would also oversee and conduct an annual audit of evaluation results carried out by trained teacher leaders, and, based on their review, would be empowered to certify district results.
5. Raise standards for judging the quality and rigor of teacher evaluation by funding and conducting studies of the reliability and validity of existing systems, including but not limited to a focus on student learning.

In summary, the authors of the Widget Effect accurately describe the apparent "educational malpractice" in the way teacher evaluation is currently implemented and used. Further, the authors conclude the report by putting forward a number of positive recommendations to move away from perfunctory, incoherent evaluation practices and toward a more comprehensive, credible and defensible system of teacher assessment. The report's primary flaw concerns its failure to incorporate and benefit from the existing body of research on teacher evaluation, as well as existing reform efforts that do move in the suggested directions. The report nevertheless stands as an important documentation of current practice in many states and districts. In the end, reinventing teacher evaluation in ways that build district capacity to strategically manage human capital will require a much greater commitment and levels of investment to support policy and practices that can both stand up to public scrutiny and be practically feasible.

## Notes and References


[1] Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The Widget Effect: Our National Failure to Acknowledge and Act on Teacher Differences*. Brooklyn, NY: The New Teacher Project. Retrieved on June 1, 2009, from http://widgeteffect.org/

[2] While the report includes Denver, which enacted a Pay for Performance reform in 2006, the report does not go into any depth to explain how the performance evaluation (based on observation) is incorporated in the pay scale. In fact, "performance" is based on several different components: 1) performance evaluation, for which a small raise is awarded for a satisfactory evaluation; 2) value-added measures of student learning; 3) willingness to work in high-need schools; and 4) degrees, National Board certification, and professional development projects. In addition, only half of Denver's teachers were paid under the new system as of May 2007 because only new teachers were required to be paid under the new system, and tenured teachers were given the choice to opt in.

[3] Rivkin, S., E. Hanushek, and J. Kain (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417-458.

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

Rockoff, J. E. (2004). The impact of individual teachers on students' achievement: Evidence from panel data. *American Economic Review 94*(2), 247-52.

[4] Goldhaber, D. and M. Hansen (2008). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research, The Urban Institute. Retrieved April 27, 2009, from http://www.urban.org/UploadedPDF/1001265_Teacher_Job_Performance.pdf.

Rothstein, J. (2008). Teacher quality in educational production: tracking, decay, and student achievement. NBER. Retrieved April 27, 2009, from http://www.nber.org/papers/w14442.

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

[5] Kennedy, M.M. (in press). Approaches to annual performance assessment. In Kennedy, M.M. (Ed.) *Handbook on Teacher Assessment and Teacher Quality*. San Francisco: Jossey-Bass. (p. 21 of chapter manuscript)

[6] For information on the CLASS instrument, see

Pianta, R. C. (2003). *Standardized classroom observations from pre-K to third grade: A mechanism for improving quality classroom experiences during the P-3 years*. Unpublished manuscript. Retrieved on March 16, 2009 from: http://www.fcd-us.org/usr_doc/StandardizedClassroomObservations.pdf.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27-50.

Mashburn, A.J., Pianta, R., Hamre, B.K., Downer, J.T., Barbarin, O., Bryant, D., Burchinal, M., Clifford, R., Early, D., Howes, C. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development, 79*(3), 732-749.

For information on the Instructional Quality Assessment (IQA), see

Junker, B., Weisberg Y., Matsumara, L.C., Crosson, A., Kim Wolf, M., Levison, A., Resnick, L. (2006). *Overview of the Instructional Quality Assessment*. (CSE Technical Report #671). Los Angeles, CA: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L.C., Slater, S.C., Junker, B., Peterson, M., Boston, M., Steele, M., & Resnick, L. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instruc-*

*tional Quality Assessment*. (CSE Technical Report #681). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Matsumura, L.C., Garnier, H., Slater, S.C., & Boston, M.B. (2008). Measuring instructional interactions 'at-scale', *Educational Assessment, 13*(4), 267-300.

[7] The assessment is based on the Danielson framework. Danielson, C. (1996/2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

See:

Gallagher, H.A. (2004). Vaughn Elementary's Innovative Teacher Evaluation System: Are Teacher Evaluation Scores Related to Growth in Student Achievement? *Peabody Journal of Education, 79*(4), 79-107.

Kimball, S. M. White, B. Milanowski, A. T. Borman, G. (2004). Examining the Relationship Between Teacher Evaluation and Student Assessment Results in Washoe County. *Peabody Journal of Education, 79*(4), 54-78.

Milanowski, A. (2004). The Relationship Between Teacher Performance Evaluation Scores and Student Achievement: Evidence From Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

[8] Wilkerson, D.J., Manatt, R.P., Rogers, M.A., Maughan, R. (2000). Validation of Student, Principal, and Self-Ratings in 360° Feedback® for Teacher Evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 179-192.

Wilcox, J.M.B. (1995). *A comparison of teachers', students' and administrators' perceptions of teaching performance quality in selected K±12 schools*. Doctoral dissertation, Iowa State University, Ames, Iowa.

[9] Luft, J. A., Roehrig, G. H., & Patterson, N. C. (2003). Contrasting landscapes: A comparison of the impact of different induction programs on beginning secondary science teachers' practices, beliefs, and experiences. *Journal of Research in Science Teaching, 40,* 77-97

[10] Smith, T. M., & Ingersoll, R. M. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal, 41*(3), 681-714.

[11] Pullin, D. (in press). Judging teachers: the law of teacher dismissals. In Kennedy, M. (Ed.). *Handbook of Teacher Assessment and Teacher Quality.* San Francisco: Jossey-Bass.

[12] Walsh-Sarnecki, P. (2007). In metro Detroit, bad teachers can go on teaching. *Detroit Free Press*. Detroit, Michigan.

[13] Honawar, V. (2007, December 5). New York City taps lawyers to weed out bad teachers. *Education Week 27(*14), 13; and Sawchuk, S. (2008). D.C. set to impose teacher-dismissal plan. *Education week 28*(10), 6.

[14] Zirkel, P. A. (2003). Legal boundaries for performance evaluation of public school professional personnel. *Education Law Reporter, 172*, 1-15.

[15] Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2007). *Who leaves? Teacher attrition and student achievement.* Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research, The Urban Institute.

Goldhaber, D., Gross, B., & Player, D. (2007). *Are public schools really losing their "best"? Assessing the career transitions of teachers and their implications for the quality of the teacher workforce.* Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research, The Urban Institute.

Hanushek, E. A., & Rivkin, S. G. (2008). *Do disadvantaged urban schools lose their best teachers?* Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research, The Urban Institute.

[16] Sykes, G., & Winchell, S. (in press). Assessing teacher tenure. In Kennedy, M. (Ed.). *Handbook on Teacher Assessment and Teacher Quality.* San Francisco: Jossey-Bass.

[17] Duffett, A, Farkas, S., Rotherham, A., & Silva, E. (2008, May). *Waiting to be won over: Teachers speak on the profession, unions, and reform.* Washington, DC: Education Sector, p. 3

[18] American Federation of Teachers/National Education Association (1998). *Peer assistance & peer review: An AFT/NEA handbook*. Washington, DC: Author.

[19] See for example

Lavely, C. (1992). Actual incidence of incompetent teachers. *Educational Research Quarterly*, 15(2), 11-14.

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

Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*(2), 103-126.

Menuey, B.P. (2005. Teachers' perceptions of professional incompetence and barriers to the dismissal process. *Journal of Personnel Evaluation in Education, 18*(4), 309-325.

[20] Marshall, R. (2008). *The case for collaborative school reform. The Toledo experience.* Washington, DC: Economic Policy Institute. See also

Gallagher, J.J., Lanier, P., & Kerchner, C.T. (1993). Toledo and Poway: Practicing peer review. In C.T. Kerchner & J.E. Koppich (Eds.), *A union of professionals: Labor relations and educational reform* (pp.158-176). New York: Teachers College Press.

Kaboolian, L., & Sutherland, P. (2005). *Evaluation of Toledo Public School District Peer Assistance and Review Plan.* Cambridge, MA: Harvard University, John F. Kennedy School of Government.

Kerchner, C.T., Koppich, J.E., & Weeres, J.G. (1997). *United mind workers: Unions and teaching in the knowledge society.* San Francisco: Jossey-Bass.

Goldstein, J. (2007). Easy to dance to: Solving the problems of teacher evaluation with peer assistance and review. *American Journal of Education, 113,* 479-508.

[21] Marshall, R. (2008). *The case for collaborative school reform. The Toledo experience.* Washington, DC: Economic Policy Institute

[22] Youngs, P., Pogodzinski, B., and Low, M. (in press). The role of formative assessments in new teacher induction. In Kennedy, M. (Ed.). *Handbook on Teacher Assessment and Teacher Quality.* San Francisco: CA: Jossey-Bass.

[23] For information on the California Beginning Teacher Support and Assessment program (both CFASST and FAS models), see

West Ed. (1997). *California Teacher Portfolio*. San Francisco: Author.

Mitchell, D.E., Scott-Hendrick, L., Parrish, T., Crowley, J., Karam, R., Boyns, D., & Mitchell, T.K. (2007). *California Beginning Teacher Support and Assessment ad Intern Alternative Certification Evaluation Study: Technical Report*. Riverside, CA: University of California-Riverside.

Glazerman, S., Senesky, S., Sefotr, N., & Johnson, A. (2006*). Design of an Impact Evaluation of Teacher Induction Programs.* Washington, DC: Mathematica Policy Research, Inc.

Glazerman, S., Dolfin, S., Bleeker, M., Johnson, A., Isenberg, E., Hugo-Gil, J. Gruder, M., Britton, E., Britton, E., & Ali, M. (2008). *Impacts of comprehensive teacher induction: Results from the first year of a randomized controlled study*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Thompson, M., Goe, L., Paek, P., & Ponte, E. (2004). *Study of the impact of the California Formative Assessment and Support System for Teachers*: Report 1, Beginning teachers' engagement with BTSA/CFASST. (CFASST Rep. No. 1, ETS RR-04-30). Princeton, NJ: Educational Testing Service.

Thompson, M., Paek, P., Goe, L., & Ponte, E. (2004). *Study of the impact of the California Formative Assessment and Support System for Teachers*: Report 3, Relationship of BTSA/CFASST engagement and student achievement. (CFASST Rep. No. 3, ETS RR-04-32). Princeton, NJ: Educational Testing Service.

For information on the Connecticut Competency Instrument and the Beginning Educator Support and Training program, see

Pecheone, R.L., & Stansbury, K. (1996). Connecting teacher assessment and school reform. *The Elementary School Journal, 97*(2), 163-177.

Wilson, S.M., Darling-Hammond, L., & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning*. Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Youngs, P. (2002). *State and district policy related to mentoring and new teacher induction in Connecticut*. New York: National Commission on Teaching and America's Future.

Youngs, P., & Bell. C. (in press). When policy instruments combine to promote coherence: An analysis of Connecticut's policies related to teacher quality. *Journal of Educational Policy.*

[24] Solmon, L.C., White, J.T., Cohen, D., and Woo, D. (2007). The effectiveness of the Teacher Advancement Program. Santa Monica, CA: National Institute for Excellence in Teaching. Retrieved July 19, 2009 from: http://www.tapsystem.org/pubs/effective_tap07_full.pdf

[25] The reviewing authors would like to acknowledge several authors whose manuscripts for a book in press (Kennedy, M. (Ed.), Handbook on Teacher Assessment and Teacher Quality) serve as the basis for and provided much of the literature for this section: Mary Kennedy on annual teacher appraisal, Diana Pullin on the law in teacher dismissal, Gary Sykes and Sarah Winchell on teacher evaluation in tenure decisions, and Peter Youngs, Ben Pogodzinski, and Mark Low on the role of formative assessments in new teacher induction. We would also like to thank Anthony Milanowski for offering his perspectives on the challenges of teacher evaluation and the reference to the Milken Family Foundation's work with the Teacher Assessment Program.

[26] See for example

Lavely, C. (1992). Actual incidence of incompetent teachers. *Educational Research Quarterly*, 15(2), 11-14.

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

Tucker, P. D. (1997). Lake Wobegon: Where all teachers are competent (or, have we come to terms with the problem of incompetent teachers?). *Journal of Personnel Evaluation in Education, 11*(2), 103-126.

Menuey, B.P. (2005). Teachers' perceptions of professional incompetence and barriers to the dismissal process. *Journal of Personnel Evaluation in Education, 18*(4), 309-325.

[27] National Council on Teacher Quality (2007). *State teacher policy yearbook: Progress on teacher quality*. Washington, DC: Author. Retrieved on July 22, 2009, from http://www.nctq.org/stpy/reports/stpy_national.pdf.

[28] Wilson, M, Hallam, 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*.

[29] See for example

Athanases, S. Z. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *Elementary School Journal, 94*(4), 421–439.

Lustick, D., & Sykes, G. (2006). National Board Certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5). Retrieved March 1, 2006, from http://epaa.asu.edu/epaa/v14n5/.

Sato, M., Chung, R., Darling-Hammond, L., & Atkin, J.M. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal*, 45(3), 669-700.

[30] See for example

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation at the University of North Carolina at Greensboro.

Cavaluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* (National Science Foundation No. REC-0107014). Alexandria, VA: The CAN Corporation.

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 117.

Smith, T., Gordon, B., Colby, S., & Wang, J. (2005). *An examination of the relationship of the depth of student learning and National Board certification status*. Office for Research on Teaching, Appalachian State University. Retrieved August 4, 2009, from http://www.nbpts.org/UserFiles/File/Applachian_State_study_D_-_Smith.pdf.

Goldhaber, D., & Anthony, E. (2005). *Can teacher quality be effectively assessed?* Seattle, WA: University of Washington and the Urban Institute.

[31] Toch, T., & Rothman, R. (2008). Rush to Judgment: Teacher Evaluation in Public Education. Washington, DC: *Education Sector*. ED502120. Retrieved on July 22, 2009, from: http://www.educationsector.org/research/research_show.htm?doc_id=656300

[32] See Wilson, M, Hallam, 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*.

Goldhaber, D., & Anthony, E. (2005). *Can teacher quality be effectively assessed?* Seattle, WA: University of Washington and the Urban Institute.

Cavaluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* (National Science Foundation No. REC-0107014). Alexandria, VA: The CAN Corporation; and

Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 117.

[33] Odden, A. & Kelly, C. (2001). *Paying teachers for what they know and do: New and smarter compensation strategies to improve schools.* Thousand Oaks, CA: Corwin Press.

The Think Tank Review Project is made possible by funding from the Great Lakes Center for Education Research and Practice.