

Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching

Anthony T. Milanowski

Consortium for Policy Research in Education

Wisconsin Center for Education Research

University of Wisconsin-Madison

1025 W. Johnson St.

Madison, WI 53706

(608) 263-4260

amilanow@wisc.edu

March 18, 2011

This paper was prepared for presentation at the American Education Research Association annual meeting on April 10, 2011 in New Orleans, LA. Its preparation was supported in part by the Ford Foundation. The opinions expressed are those of the author and do not necessarily reflect the view of the Ford Foundation or the Wisconsin Center for Education Research.

The contributions to this research of current and former colleagues Steven Kimball, Herbert Heneman, Allan Odden, H. Alix Gallagher and Bradford White, are gratefully acknowledged.

Validity Research on Teacher Evaluation Systems Based on the Framework for Teaching

After decades of disinterest, evaluation of the performance of elementary and secondary teachers in the United States has become an important educational policy issue. The US Federal government has made teacher evaluation a focus of its Race to the Top grant program for states, and part of its Teacher Incentive Fund grant program for states and districts. To receive funds, potential grantees must show they are taking performance evaluation seriously as a strategy to improve the effectiveness of the teacher workforce and to hold teachers accountable for performance. Several large US cities, including New York, Chicago, and Washington DC have also been trying to use teacher evaluation as a strategy to improve instruction, often beginning by identifying weak teachers for remediation or termination.

At the same time, reformers from outside the government have also made teacher evaluation an issue. For example, the New Teacher Project's report, *The Widget Effect*, (Weisberg, Sexton, Mulhern, & Keeling, 2009) documented for the public the fact well known to both district leaders and researchers that performance evaluation ratings fail to differentiate much among teachers, with few rated in the lower categories (see also Dwyer & Stufflebeam, 1996; Loup, Garland, Ellett, & Rugutt, 1996). This report has received a lot of attention in policy circles, and along with other critiques of what has been taken as typical practice¹ and suggestions for improvement by reformers (e.g., Toch and Rothman, 2008; New Teacher Project, 2010).

¹ The conventional description of teacher evaluation in the US has been that a principal or assistant principal makes one classroom observation per year and either fills out a checklist or takes note and composes a brief narrative, then makes a final overall rating of satisfactory or unsatisfactory that goes into a teacher's file and, if satisfactory, is then forgotten.

More rigorous assessment of teaching has also been promoted by the movement to “professionalize” teaching in the US. The National Board for Professional Teaching Standards certification assessments, the Educational Testing Service’s PRAXIS III observation-based assessment for new teacher licensure, and the Interstate New Teacher Assessment and Support Consortium’s standards for teacher licensure are results of this movement. Among the common themes in these assessments are a view of teaching as a complex activity not well captured by checklists and brief observations, and requiring the application of professional judgment to assess (Porter, Youngs and Odden, 2001). Another product of the professionalization movement has been Charlotte Danielson’s Framework for Teaching (Danielson, 1996). Since its introduction, the Framework for Teaching has, by merit and by default, become part of the foundation for efforts to improve teacher evaluation in the US.

Interest in teacher evaluation has also been fed by the development of statistical methods to estimate the effects of individual teachers on student learning (such as value-added modeling).² Beginning in the late 1990’s, studies using value-added methods provided evidence of substantial variation in teachers’ contributions to student learning (e.g., Sanders and Rivers, 1996; Wright, Horn, and Sanders, 1997; Webster, Mendro, Orsak, and Weerasinghe, 1998; Rowan, Correnti, and Miller, 2002). These studies showed much more variation in teacher effectiveness than reflected in the rating of 99% of teachers as satisfactory performers and 1% as unsatisfactory, per common UIS teacher evaluation practice.

Not only has value-added shown how varied teacher effects on student achievement are, but this technology seems both simpler and more precise than something as seemingly subjective as observing teaching practice. Proposals have been made to base consequential decisions such

² See Harris, 2011 for a good non-technical discussion of value-added modeling and its implications.

as termination solely on value-added estimates of teacher effectiveness (e.g., Gordon, Kane, and Staiger, 2006). The US Federal government has emphasized giving value-added substantial weight in teachers' evaluation results as part of its Teacher Incentive Fund and Race to the Top grant programs. The need to find teaching practice measures with more rigor to compliment value-added for use in performance pay systems, for tenure decisions, and to identify professional development needs has led many US states and school districts to look for a better evaluation model.

One of the models that has been attracting attention is the Framework for Teaching (Danielson 1996, 2007). The Framework is being used as a basis for teacher evaluation by a substantial number of US school districts. While no census has been done, based on personal familiarity I estimate that at least 200 US school districts use the Framework or variations. There will likely be many more since at least two US states (Delaware, Idaho) are poised to use it as the basis for a statewide teacher evaluation model.

This paper attempts to summarize validity evidence pertaining to several different implementations of the Framework. It is based primarily on reviewing the few published and unpublished studies that have looked at the relationship between teacher evaluation ratings made using systems based on the Framework and value-added measures of teacher effectiveness. However, much as states or districts want evidence that the Framework is the "right" model to use, the evidence we have is not about the validity of the Framework itself. The evidence is about the validity of the inference made based on the evaluation ratings in specific Framework implementations. The way the Framework is used – the procedures for training evaluators, collecting evidence, combining that evidence, and making a rating decision - are likely to be as influential in decisions about individual teachers' practice as the content of the Framework's

components, elements, and rubrics. An evaluation system is indeed a system, and the procedural variations among different implementations of the Framework likely have a lot to do with differences in the reliability or validity of ratings. The question then is not whether the Framework is valid, but rather whether ratings made using evaluation systems based on it are validly used to infer something about the quality of the rated teachers' practice or performance, including that teachers who receive different ratings have different levels of teaching performance. Nevertheless, the evidence reviewed in this paper does have value because it shows that evaluation systems based on the Framework can produce reliable ratings that correlate with value-added estimates of teachers' contributions to student achievement.

Overview of the Framework for Teaching

The Framework for Teaching (hereafter, FFT) is a set of teaching performance dimensions or standards accompanied by a set of rating scales or rubrics, each describing four levels of performance (Danielson, 1996, 2007). These levels are labeled unsatisfactory, beginning, proficient, and distinguished. The FFT is intended for use at the elementary and secondary levels and be applicable across subject areas. It partitions teaching into four "domains": planning and preparation, the classroom environment, instruction, and professional responsibilities. Each domain has specific performance "components" and then "elements" nested within them. Each element has a rubric or rating scale that describes the four performance levels in terms of observable teacher or student behavior. The domains and components of the FFT are shown in Table 1.

Table 1

Domains & Components of the Framework for Teaching

| Domain | Components |
|----------------------------------|---|
| 1. Planning and Preparation | 1a: Demonstrating Knowledge of Content and Pedagogy 1b: Demonstrating Knowledge of Students 1c: Setting Instructional Outcomes 1d: Demonstrating Knowledge of Resources 1e: Designing Coherent Instruction 1f: Designing Student Assessments |
| 2. The Classroom Environment | 2a: Creating an Environment of Respect and Rapport 2b: Establishing a Culture for Learning 2c: Managing Classroom Procedures 2d: Managing Student Behavior 2e: Organizing Physical Space |
| 3. Instruction | 3a: Communicating with Students 3b: Using Questioning and Discussion Techniques 3c: Engaging Students in Learning 3d: Using Assessment in Instruction 3e: Demonstrating Flexibility and Responsiveness |
| 4. Professional Responsibilities | 4a: Reflecting on Teaching 4b: Maintaining Accurate Records 4c: Communicating with Families 4d: Participating in a Professional Community 4e: Growing and Developing Professionally 4f: Showing Professionalism |

Note: Adapted from the Framework for Teaching (Danielson, 1996)

The FFT was originally developed as a basis for conversations between teachers and administrators about teaching practice. The idea was to provide a common language about teaching that would incorporate current view of effective practice. The FFT has its origins in work Danielson did on the PRAXIS III licensure assessment. PRAXIS III was intended measure the classroom practice of new teachers, as a supplement other the measures of teaching skills in making a licensing decision. The FFT built on PRAXIS III by including more aspects of teaching and defining higher (more accomplished) levels of practice. After the FFT was published, several US school districts looking to improve their teacher evaluation systems began to use the FFT for

summative evaluations (see Kimball, 2002). Because the FFT’s content recognized the complexity of teaching and provided ready-made behaviorally-referenced rating scales, it was an attractive replacement for checklists of behaviors or unstructured narratives as evaluation tools.

Danielson (1996, 2007) encouraged modification of the FFT to better fit the local context. Districts appear to have done this to a greater or lesser degree. Of the districts my colleagues and I have studied, most made only minor modifications. One, the Cincinnati Public Schools, made extensive modifications. It reduced the 22 components of the FFT to 15 standards. Table 2 shows two elements of the original FFT compared with the Cincinnati versions that were derived from them.

Table 2
Comparison of Selected Content of the Framework for Teaching and Cincinnati’s Version

| Framework for Teaching (2007) “Distinguished” Level | Cincinnati “Distinguished” Level |
|---|---|
| <p>Component 3 b Using Question and Discussion Techniques. Element: Quality of Questions</p> <p>Teacher’s questions are of uniformly high quality with adequate time for students to respond. Students formulate many questions.</p> | <p>Standard 3.4 Part B: Thought- Provoking Questions</p> <p>Teacher routinely asks thought - provoking questions at the evaluative, synthesis, and/or analysis levels that focus on the objectives of the lesson. Teacher seeks clarification and elaboration through additional questions. Teacher provides appropriate wait time.</p> |
| <p>Component 3d Using Assessment in Instruction, Element: Assessment Criteria</p> <p>Students are fully aware of the criteria and performance standards by which their work will be evaluated and have contributed to the development of the criteria</p> | <p>Standard 3.1 Part E: Assessment Criteria</p> <p>Teacher clearly communicates assessment criteria that are aligned with the standards-based instructional objectives and includes the task-specific criteria for various performance levels.</p> |

As originally presented by Danielson, the FFT suggests, but does not specify, procedures for observing teaching and coming up with a rating. Districts using it have used a variety of procedures for gathering evidence and making evaluation ratings. For example, in several of the

districts we have studied, school administrators are the sole observers and raters, and observations of veteran teachers are made one or two times per year. Some require teachers to select and share a specified set of artifacts such as lesson plans and student assignments, while others make this voluntary. An example of one of the more rigorous procedures is that of the Cincinnati Public Schools. For a comprehensive evaluation, which during the last several years has been done for veteran teachers every 5 years after the third year, five (later four) classroom observations were made, three by observers from outside the school and two by school administrators. One observation by each was unannounced, so that teachers cannot prepare special lessons in advance. Beginning teachers were observed every year, with 5-6 observations made by evaluator/mentors from outside the school. Observers participated in a multi-day training session, and had to achieve a set level of agreement with an expert panel of observers when rating a set of videotaped lessons, in order to be allowed to evaluate. The final rating was determined by the preponderance of ratings from the observations, plus assessments of artifacts such as lesson plans and teachers' reflections on the observed lesson.

Validity Evidence for Evaluation Systems Based on the FFT.

There are two separate but related senses of “validity” that are of interest to US policy makers. First, they tend to want to know whether the teaching practices described in a model like the FFT lead to higher levels of student learning. This is what they want to know when they ask “Is the FFT valid?” The psychometrically sophisticated cringe at this question, preferring to think of validity as a property of the inference that teachers rated high or low using an evaluation process based on the FFT are in fact better or worse teachers. Of course, policy makers should also be concerned with the second sense if they want to use evaluation ratings for consequential decisions, But the first sense is also important, because the definition of better or worse teacher is typically based heavily on the concept of student learning.

Validity studies of FFT-based evaluation systems have primarily focused on the relationship between teachers’ evaluation ratings and teachers’ effects on student learning, as represented by classroom-level value-added estimates of teacher productivity. This type of evidence used to be called *criterion-related validity* evidence, based on the idea that if there is an external standard of performance (the criterion) then ratings should correlate with or predict measures of the standard. For many policy makers and educational leaders, value-added is the accepted criterion, if not definition, of teacher effectiveness, so researchers have focused on the relationship of these estimates with evaluation ratings. While this is not the only type of validity evidence that matters, it has been the most commonly sought, because of the interest in finding a measure of practices that facilitate student learning. This type of evidence is also relevant to the question of whether inferences based on evaluation scores are justified, given the assumption that the criterion represents a valued outcome.

Criterion-Related Validity Studies

Some evidence about the relationship of evaluation ratings from FFT-based systems and value-added comes from research I have done with colleagues Steve Kimball, Herbert Heneman, Brad White, Alix Gallagher, and Allan Odden. We studied the relationship between classroom value-added estimates for reading, math, and at one site science, and evaluation ratings for teachers in elementary and sometimes middle grades using systems based relatively closely on the FFT. We obtained data from three sites in the US: the school districts of Cincinnati, Ohio, Coventry, Rhode Island, and Washoe County, Nevada.

Table 3 shows the correlations between teachers' evaluation scores and value-added estimated at the classroom-level at three of these sites. In the table, the correlations are between total or average evaluation scores and a value-added estimate based on empirical Bayes intercept residuals from a two-level hierarchical linear model. Only teachers with test scores from three or more students in any year were included. In order to summarize the relationship between teacher evaluation scores and student achievement across years for each site, we combined correlations for grades within subjects to produce a single estimate for each subject and year within sites. Each grade within a subject is treated as a separate study and the correlations combined using the meta-analysis formulas for a random effects treatment.³ Upper and lower bounds for the 95% confidence intervals were also calculated. We then took the weighted average of the correlations across years within subjects and sites.

³ An r to z transformation was done and a weighted average of the z 's was calculated with the inverse of the variances as weights. Standard errors were calculated for this average, and 95% confidence intervals. These values were then transformed back into correlation coefficients.

Table 3

Correlations Between Teacher Evaluation Scores and Estimates of Classroom Value-Added

| Site | Grades | Tested Subject | | |
|----------------------|---------|----------------|-------------|----------------|
| | | <u>Reading</u> | <u>Math</u> | <u>Science</u> |
| Cincinnati | | | | |
| 2001-02 | 3-8 | .48* | .41* | .26* |
| 2002-03 | 3-8 | .28* | .34* (4-8) | -.02 (4-8) |
| 2003-04 | 3-8 | .29* | .22 | .29* |
| 3 year wtd. average: | | .35 | .33 | .18 |
| Coventry | | | | |
| 1999-2000 | 2,3,6 | .17 | .05 | |
| 2000-2001 | 2,3,4,6 | .24 | -.17 | |
| 2001-2002 | 4 | .39 | .34 | |
| 3 year wtd. average: | | .24 | -.06 | |
| Washoe | | | | |
| 2001-02 | 3-5 | .22* | .20* | |
| 2002-03 | 4-6 | .25* | .24* | |
| 2003-04 | 3-6 | .19* | .21* | |
| 3 year wtd. average: | | .22 | .22 | |

* 95% Confidence interval does not include 0.

In general, the correlations are positive, but vary across sites and across years. The Washoe correlations are the most stable, which might be expected given the large sample sizes, and because many of the same teachers are included in the analyses across years. The negative correlation for math in Coventry for 2000-01 surprised us, but the sample size in this district was quite small. The small negative correlation in Cincinnati for science in 2002-03 is also anomalous. Other than these, if we consider the correlations as effect size measures, we might be characterize them as between small ($r=.1$) and medium ($r=.3$).

If teacher evaluation scores represent teacher expertise, rather than being due to transitory factors in the evaluation year, we would expect to see positive correlations between teacher evaluation scores and classroom value-added from the year after evaluation. Table 5 shows the correlations between classroom value-added and the teacher evaluation scores from two of our sites.

Table 5

Correlations Between Student Achievement and Teacher Evaluation Scores from the Prior Year

| Site | Tested Subject | | |
|--|----------------|-------------|----------------|
| | <u>Reading</u> | <u>Math</u> | <u>Science</u> |
| Cincinnati | | | |
| Teachers Rated in 01-02, Student Achievement from 02-03 | .20* n=109 | .20 n=76 | .13 n=35 |
| Teachers Rated in 02-03 Student Achievement from 03-04 | .20 n=61 | .21 n=36 | .19 n=32 |
| Washoe | | | |
| Teachers Rated in 01-02 Student Achievement from 02-03 (n=248) | .11 | .19* | - |
| Teachers Rated in 02-03 Student Achievement from 03-04 (n=229) | .14* | .04 | - |

* 95% confidence interval does not include 0
Source: Milanowski & Kimball, 2005.

Complementary evidence has been provided by another study of the Cincinnati system by Kane, Taylor, Tyler, and Wooten (2010). This study used more years of teacher evaluation and student test score data, but because it used a very different analytical strategy, it is difficult to compare the results directly with those in Tables 4 and 5. Kane et al also used only ratings from

standards in the classroom environment and instruction domains, and used the difference between actual and predicted student achievement from the year *after* the teacher was evaluated rather than in the same year.³ Based on a principal components analysis of ratings on eight standards in these two domains, they constructed three evaluation scores. The first was the average of the eight standard ratings (interpreted as overall performance), the second, the difference between the averages of the ratings on the Domain 2 and 3 standards (interpreted as an emphasis on versus classroom management) and the third, the difference between the average on standards related to the more routine aspects of teaching and ratings on the standard related to questioning (interpreted as a contrast between teaching through questioning and discussion and focusing on classroom management and conveying standards-based instructional objectives).

While there are several interesting sets of results presented in this study, only the broader findings will be discussed here. First, teachers in the top value-added quartile consistently received higher ratings on all the standards in the prior year than those in the bottom. Second, a one point increase in the average of evaluation ratings on the eight standards is associated with a one-sixth standard deviation increase in math achievement and a one-fifth standard deviation increase in reading achievement, controlling for the other two evaluation scores the researchers constructed and the typical control used in a value-added model. Given that this one point increase is equal to two standard deviations in teacher average rating, the difference in achievement for a teacher with an average rating one standard deviation higher is .08 of a standard deviation in math and .10 of a standard deviation in reading. A teacher whose average score would place her at the ‘distinguished’ level will have students whose student achievement is about one-fifth of a standard deviation higher in that a teacher at the ‘proficient’ level. Slightly

³ This was done to eliminate the effect of unobserved student characteristics that might contemporaneously affect both the evaluation rating and student achievement. See pages 11-13 of the Kane et al paper.

larger effects were found for the average of the eight standard ratings when using student achievement from the same year as the observations were made. The differences were not statistically significant, however. Interestingly, the average evaluation score measure was positively and significantly related to student achievement in the prior year as well.

A positive effect was found on the Domain 2 minus 3 measure in reading and math but this was much smaller and not significant in the model that used student achievement data from the observation year. Kane et al interpret this result as showing that controlling for overall teacher performance, those teachers who emphasize classroom management get better results. They speculate that “Cincinnati may be operating in the range of the education production function where increases in classroom environment inputs such as keeping kids on task have bigger payoffs to student achievement than increases to inputs associated with instructional practices ...” (page 21).

When interpreting the Kane et al results as evidence of validity of inferences about teachers from their evaluation scores, it should be noted that the teaching measures developed by these researchers were calculated in a way quite different from that used by the district, and do not include the domain 1 and 4 ratings, as the district does. Nevertheless, these results do show that FFT-based evaluation ratings can have a substantial relationship with student achievement.

Criterion-Related Studies of Systems Based on Framework “Cousins”

As mentioned above, Danielson has encouraged modifications of the Framework. We have some results showing the relationship between value-added and evaluation ratings from two systems that represent more radical modifications of the FFT’s rubrics than we saw in Cincinnati. First, with former colleague Alix Gallagher, I examined this relationship at the Vaughn Learning Center, a charter school in Los Angeles. For the first three years of its

performance pay system, this school used selected components from Domains 1 and 2, but replaced Domain 3 (Instruction) with subject-specific rubrics for literacy, math, and language arts instruction, as well as one for special education inclusion. (In the fourth year, the school switched to a system based more closely on the original FFT.) In this system, an administrator, a peer teacher, and the teacher being evaluated all rated performance, and the ratings were averaged. Peers and administrators observed multiple times during a two week period each semester. Table 5 shows correlations between classroom value-added estimates and evaluation ratings from the Vaughn Learning Center. The average number of teachers covered in each year was 30.

Table 5
Correlations Between Teacher Evaluation Scores and Estimates of Classroom Value-Added for Vaughn Learning Center and TAP

| Site | Grades | Tested Subject | | |
|-------------------------------|--------|----------------|---------------------------|---------------|
| | | Reading | Math | Language Arts |
| Vaughn | | | | |
| 2000-01 ^a | 2-5 | .48* | .20 | .20 |
| 2001-02 ^b | 2-5 | .58* | .42* | .42* |
| 2002-03 ^b | 2-5 | .05 | .17 | .29 |
| 3 Year Average: | | .37 | .26 | .30 |
| NIET-TAP Study 1 ^c | 4-6 | .68* | .55* | .70* |
| NIET TAP Study 2 ^d | ? | | .21 (subjects not stated) | |

- a) From Gallagher, 2004
- b) From Milanowski & Kimball, 2005
- c) From Schacter & Thum, 2004
- d) From Daley & Kym, 2010

Another evaluation system related to the FFT is that used by the National Institute for Excellence in Teaching's TAP school improvement model. The TAP model includes a teacher

evaluation system with rubrics that were based in part on the FFT. The system's four domains, designing and planning instruction, learning environment, instruction, and professional responsibilities roughly parallel the FFT, but the rubrics are tailored to the aspects of instruction that NIET believes are critical for improving student learning. Notably, the rubrics for the instruction domain are more specific about practices relating to student engagement, teaching to standards, higher order thinking skills, use of assessment, and differentiation of instruction. The TAP evaluation system includes multiple observations by both school administrators and teacher leaders (a school master teacher and team leaders called mentor teachers) and the results are used as one factor in determining teacher bonuses. Two studies containing evidence on teacher rating – value-added correlations are available. First, a study by Schlacter and Thum (2004) used evaluations made of 52 teachers by trained outside evaluators (graduate students) using an early version of the TAP evaluation system. These evaluators made eight observations of each teacher. The study found correlations of ratings with classroom value-added of .55 to .70 (See Table 5 above). More recently, a study by the NIET's Daley and Kim (2010) used a larger sample of 1,780 teachers. The study found that there was a significant coefficient for the evaluation rating in a regression of transformed classroom value-added estimates on evaluation ratings. As reported in the study, the coefficient of determination for this model was .452, which corresponds to a correlation coefficient of .21, which can be more readily compared to our results for the Framework.⁴

While the samples at Vaughn and TAP Study 1 are small, it is interesting that the correlations shown in Table 5 are higher than those in Table 3. Could it be that the more specific rubrics of

⁴ Note that the relationship reported by Daley & Kym is likely to be understated because the value-added indicators used were the result of grouping the original value-added estimates into 5 score categories. This reduces the amount of variance available for explanation by the evaluation ratings.

the Vaughn and TAP systems help evaluators do a better job of distinguishing among teachers on practices that matter to learning? Or is it just that rater training was better or the total number of observations greater? Another interesting feature of Table 6 is the drop in correlations at Vaughn in the third year of our study. I will return to this in the concluding section.

Comparable Evidence from Other Studies

To better judge the evidentiary value of these correlations it would be useful to have more evidence on the relationship between teacher evaluation scores from a wider range of evaluation systems. Unfortunately, there is not yet very much published research with which to compare the results reviewed above. I have located only two studies that reported comparable student achievement-value-added associations. Fritsche et al (2002) found correlations of .17 and .24 between evaluation scores based on the Texas state evaluation system and gains in student achievement. Jacob and Lefgren (2005) found correlations between reading and mathematics value added and principals' rating of how effective a teacher is at raising student achievement of .20 and .28 respectively. These ratings were guided by a much simpler rubric that did not explicitly describe practice, and were made without having to be shared with teachers.⁵ These few studies seem to suggest that the Framework-based systems can produce ratings with comparable or stronger relationships with value-added indicators of teacher effectiveness.

There will soon be substantially more evidence of the relationship of evaluation scores to value-added. The Gates Foundation's Measuring Effective Teaching project will be reporting on an extensive study of how well ratings of teaching practice based on the FFT –and other rubrics- made from videotapes of teaching practice by trained raters predict classroom value-added. This

⁵ These correlations are likely to be higher than in situations in which ratings would have to be shared with teachers, because when ratings are shared, evaluations tend to be more lenient, in turn reducing the variance of the ratings and lowering the potential correlation.

study includes 2,000 volunteer teachers in six school districts. Teachers in grades 4-8 plus, in some states, those teaching Algebra 1, 9th grade English, and biology are participating, making this the largest and most comprehensive study of its kind ever done in the US. (See Bill & Melinda Gates Foundation, 2010, for a description of the study and preliminary results).

Taken together, the research results summarized above suggest that teacher ratings made using evaluation systems based on or related to the Framework for Teaching can have a substantial relationship to value-added estimates of teacher effectiveness. While it might seem that a correlation of .2 to .3 is low, there are many factors working to attenuate the relationship between value-added and evaluation ratings. These include measurement error in both evaluation ratings and student test scores, imperfect alignment of the curriculum with tests, and misalignment between the pre- and posttests used in the value-added analyses all attenuate any positive relationship. Recent research (e.g., Schochet & Chiang, 2010) suggests that there is considerable measurement error in classroom-level value-added estimates based on just a pre- and post-test. Finally, since student learning is co-produced by teacher, student, classroom peers, and family and depends in part on student effort that teachers may influence but cannot control, a very high correlation would actually be suspicious.

Reliability Evidence

Evidence for reliability is relevant to the question of validity of inference that ratings represent actual teaching performance rather than to the question of whether teacher performance leads to student learning. Both teachers and other users of evaluation ratings want to be sure that the rating is mostly due to observed teacher performance rather than raters' idiosyncratic views about what good performance looks like or the time performance was observed. Forms of reliability such as inter-rater agreement and inter-temporal agreement are typically the priority,

in part to provide teachers with evidence of freedom from idiosyncratic rater or time effects when evaluation results are to be used for decisions with consequences. Unfortunately, there is less evidence of these types of agreement for evaluation ratings based on the FFT.

We were able to obtain estimates of rater agreement for two of our research sites: Cincinnati, Ohio, and the Vaughn Learning Center. Agreement estimates were also obtained for ratings made using an FFT-based evaluation system in the Chicago Public Schools from a report by the Consortium on Chicago School Research (Sartain, Stoelinga, and Brown, 2009). This evidence is summarized in Table 6.

Table 6
Rater Agreement Evidence from Three Framework for Teaching Sites⁶

| <u>Site</u> | <u>Inter-Rater Agreement</u> | <u>Notes</u> |
|--|------------------------------|--|
| Chicago Public Schools ^a | | Average % absolute agreement between 2 raters on components of Domains 2 & 3; n=277 teachers. |
| Domain 2 | 54% | |
| Domain 3 | 52% | |
| Cincinnati Public Schools ^b | | Average % absolute agreement on Domain 2 & 3 scores between two raters rating on different occasions; n = 99 teachers. |
| Domain 2 | 73% | |
| Domain 3 | 79% | |
| Vaughn Learning Center ^c | | Average correlation between administrator and peer raters on planning (components of Domain 1) & classroom management (components of Domain 2); average n = 48.6 teachers. |
| Domain 1 | .65 | |
| Domain 2 | .70 | |

Sources:

- a) Sartain, Stoelinga, & Brown, 2009
- b) Heneman & Milanowski, 2003
- c) Calculated by author from unpublished data

⁶ Further details about the Cincinnati estimates can be found in the appendix.

Inter-rater agreement is certainly better than chance, but it would be more reassuring to see it higher. Sartain, Stoelinga, and Brown (2009) show that in general school administrators were more lenient than external raters. We saw evidence of administrator leniency in Cincinnati and Vaughn as well. This is one plausible reason for inter-rater disagreement.

Given the likelihood of leniency, it would seem advisable to have more than one rater observe and then average the ratings. Sartain, Stoelinga, and Brown (2009) report a reliability of .94 based on multi-facet Rasch analysis of aggregated component level ratings from two raters. This corresponds to the reliability of a total score based on two observers looking at the same two classroom segments. Note that the procedure used controlled for rater severity. However, during the normal operation of an evaluation system, two raters will rarely observe on the same occasion. At the Vaughn site, we were able to examine the reliability of domain scores based on averaging the peer and administrators' ratings based on different occasions of observation during the same two week window. The alpha reliability was .78 for planning and .82 for classroom management. Using unpublished Cincinnati data, I calculated a Spearman-Brown estimate of reliability for the standards within Domain 2 and 3 based on 2 observations made by 2 raters rating on different occasions. The average across the three Domain 2 standards was .61, and for the 6 Domain 3 standards, .51. Clearly, either more observers or more occasions of observation are needed for a reliable average. Luckily, Cincinnati had chosen to require more occasions of observation.

There are good reasons to observe on more than one occasion. At our sites, we heard teachers raise the issue of variation in teaching behavior from occasion to occasion: "they saw me on a bad day". This concern is supported by some research (e.g., Rowan, Harrison, & Hayes, 2004; Rogosa, Floden, & Willett, 1984) which has shown that teaching practice does vary over

the day, week, and year. We have some evidence of inter-temporal agreement evidence from our Vaughn site. Administrators' ratings correlated .92 for planning and .83 for classroom management across semesters, while peer correlations were .79 and .68. The alpha reliabilities across semesters but within rater types were in the .80 to .95 range. Using the Cincinnati data, five observations by the same evaluator would provide an average Spearman-Brown reliability of .80 for the three standards in Domain 2, though six would be needed for the six standards in Domain 3. (see the Appendix for more details.)

Though fragmentary, the results discussed above do suggest that evaluation ratings from systems based on the FFT can show substantial inter-rater agreement and, if multiple observers and multiple occasions of observation are used, the ratings can be quite reliable. It does not seem that a score based on one observation by one rater would have acceptable reliability, and even an average of two raters' ratings of one occasion may often not be reliable enough. When evaluation results will be used for high stakes, it would be advisable to have results from generalizability studies conducted in as near to operational conditions as possible to obtain evidence on the contribution of raters and occasions of observation to rating variance. If generalizability was high, we would have more confidence that aspects of the evaluation system that are hard to control -the rater and when practice is observed- do not dominate the results.

Conducting Validity Research on Teacher Evaluation Systems

This section makes two arguments. First, that further pursuit of what I have been calling criterion-related validity evidence needs to be done in a more sophisticated way. Second, validity research needs to be expanded to include more proximal evidence of construct validity.

Pursuing Better Criterion-Related Validity Evidence

At least in the US, the relationship between ratings made using teacher evaluation processes and value-added measures of productivity or effectiveness are likely to continue to be

of interest to policy makers. However, this should not stop researchers from doing more sophisticated analyses of this relationship. Below are four suggestions for improving this line of research.

First, the likelihood that some of the evaluation dimensions may not have a strong direct linear effect on student achievement needs to be considered. Some behavior dimensions may have thresholds for facilitating a minimal level of learning, while others are needed to move from average to high levels. Consider classroom management and differentiation of instruction. Teachers need to exhibit a certain threshold level of classroom management to facilitate learning, but after that level is reached, additional increments may not contribute much more. For example, a rating of *proficient* for classroom management performance may distinguish teachers whose students demonstrate average achievement from those whose students demonstrate poor achievement, but a rating of *outstanding* may not be associated with yet higher student achievement. In contrast, a rating of *outstanding* on differentiation of instruction may distinguish teachers whose students show above-average achievement from those whose students show average achievement. In this example, a one-level difference in ratings (from *proficient* to *outstanding*) does not have the same impact on achievement across performance dimensions. It may even be that the two dimensions have an interactive relationship, with high levels of differentiation of instruction being effective only when combined with a threshold level of classroom management.

We saw some evidence of a non-linear relationship between classroom management and value-added at our Vaughn Learning Center site, as shown in Figure 1.

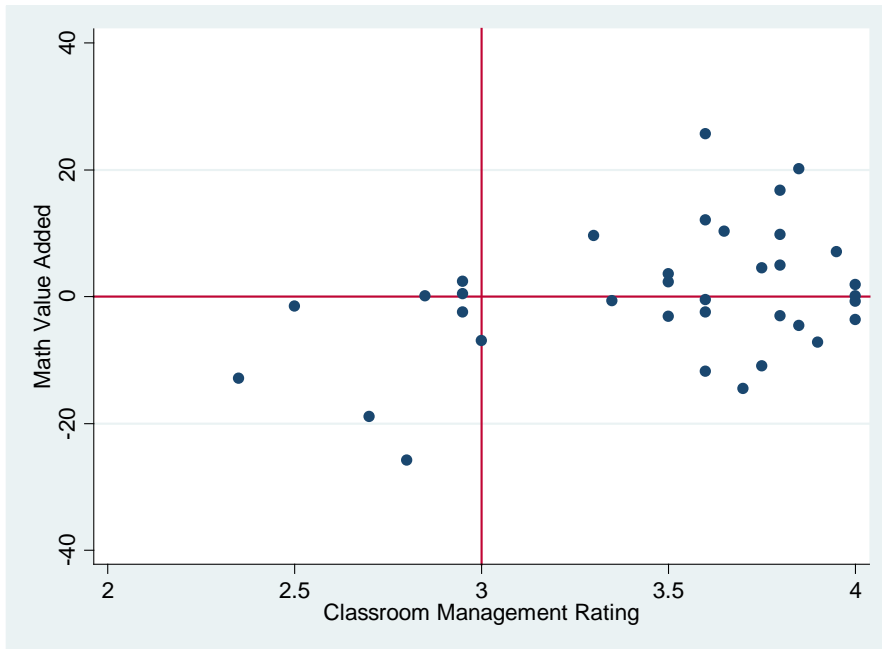


Figure 1
Relationship of Classroom Planning Rating to Math Value Added, 2001-02

Here, it seems that teachers who score above “proficient” (3) on classroom management have higher value-added, but there is little relationship above that level. This suggests that it may be appropriate to disaggregate ratings on the various dimensions of teaching performance and analyze the relationship between ratings on each dimension and the criterion measure. We might find that the overall correlation between value-added and classroom management is relatively low, but at low levels of classroom management there is a clear relationship. Similarly, there might be no overall relationship between constructivist pedagogy and value-added, but some for relationship for teachers exhibiting higher levels of that practice. Kane et al’s (2010) finding that teachers who emphasize classroom management, controlling for average classroom management and instruction ratings, also suggests that practice dimensions may have more than independent effects.

Second, just as non-random assignment of students to teachers can bias value-added measures of effectiveness, it is also likely to affect teacher evaluation ratings. For example, a class dominated by conscientious students with good prior-year preparation should make it easy to teach in ways that deserve high FFT ratings (e.g., on the Engaging Students in Learning, Activities and Assignments Element: “All students are cognitively engaged in the activities and assignments in their exploration of content. Students initiate or adapt activities and projects to enhance their understanding.”) Such a class is also likely to have high average value-added. In this case the composition of the class may be the cause of the teacher’s high rating and the high average value-added. Just as we would be more comfortable interpreting value-added as a measure of the teachers’ effect were students randomly assigned to classrooms, we would also be more comfortable with evaluation scores as indicators of teachers’ “true” performance were students randomly assigned to teachers. Though this will rarely be done in practice, for the purposes of validity research random assignment would be highly desirable. Here the Gates Foundations MET project is leading the way with random assignment of students to teachers within schools.

Third, validity studies might also make use of multiple years of performance measures. As mentioned above, recent research has made it clear that value-added estimates contain substantial sampling error. While this is much less of a problem in validity research than in constructing a point estimate of teacher effectiveness for use in consequential decisions, it still attenuates the relationship between evaluation ratings and value-added estimates. One way to reduce error would be to combine multiple years of value-added estimates. In a validation research study, one could combine multiple years of both value added estimates *and* teacher evaluation ratings. Such measures should be more stable and better represent teachers’ ability to

teach according to the model underlying the evaluation system and the effects of this ability on student achievement. A simple confirmatory factor analysis, as illustrated in Figure 2, would provide a good way of combining performance measures across years.

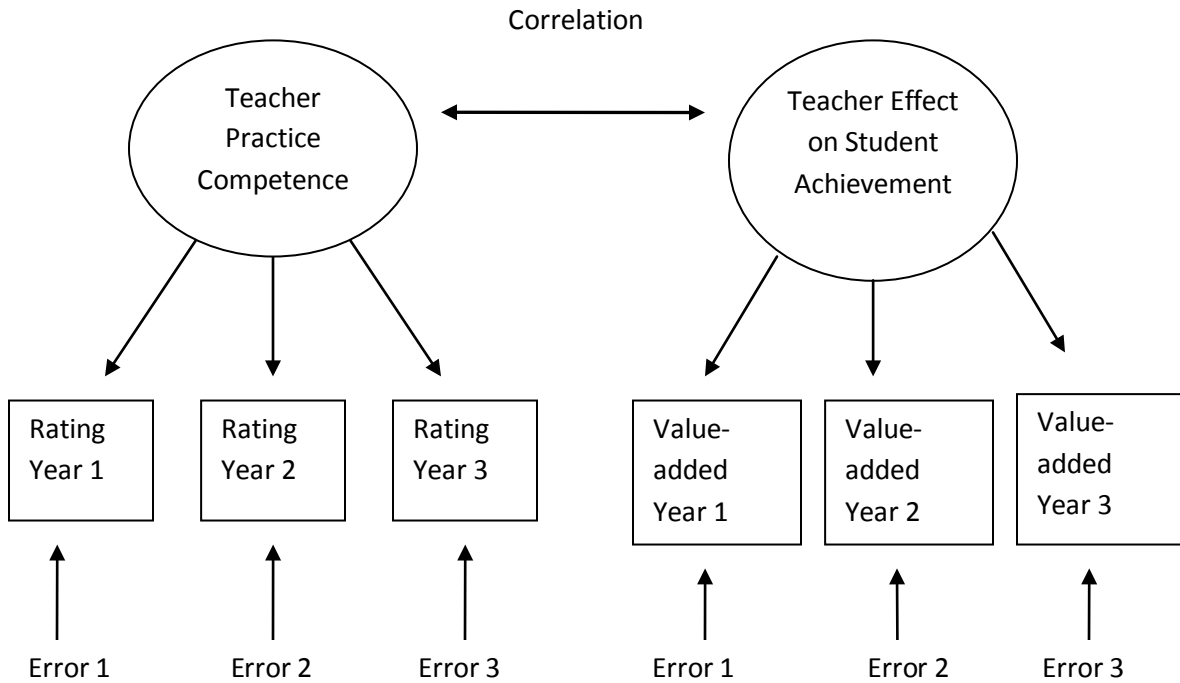


Figure 2

Confirmatory Factor Analysis Model for Criterion-Related Validity Evidence

One drawback of this approach is that we are no longer looking at evidence for the validity of inference from teacher evaluation ratings for a particular school year and teacher performance in that year. Rather, we are positing some underlying practice competence that causes the stable effect on student learning represented by three years of value-added estimates. This is not a problem if our interest is in evidence that the practices assessed by the evaluation system facilitate student learning, or if our evaluation system combines three years of ratings as the basis

for consequential decisions. But if we are interested in the validity of yearly ratings, this design is less useful.

Fourth, if evaluation systems work as intended, ratings will over time show weaker relationships with value-added estimates of teacher effects on student achievement. Two forces are at work. First, if teachers use the feedback from their evaluations to improve their practice, over time more teachers will be rated at the higher levels, and variation in ratings may decline. Lower variation in ratings lowers the potential correlation they can have with value-added measures of teacher effects, even though the evaluation process has not changed. The second factor is a reduction in the variation in value-added across classrooms. If teachers improve their practices to get higher evaluation scores, and the evaluation systems reliably measure the teaching practices that contribute to higher levels of student achievement, student achievement can go up but the variation in the teacher effect measures will go down. Again, limited variation means a limit on the size of any correlation between the evaluation ratings and the measure of teachers' effect on student achievement. We actually saw some of this effect over time at the Vaughn Learning Center, where we had value-added and evaluation measures for many of the teachers over three or four years. The school was actively pursuing a relatively prescriptive approach to reading instruction, and as teachers' reading instruction converged around the model underlying that dimension of the evaluation system, reliable variation across classrooms in reading value-added declined to near zero, and the correlation between ratings and value-added dropped to nearly zero.

Pursuing Construct Validity Evidence

From a psychometric perspective, it is clear that the search for validity evidence should not be limited to analyzing the relationship between evaluation ratings and value-added estimates. Evaluation systems based on the FFT are intended to measure teaching practice not

student learning. Teaching practice and student learning have different sets of causes and we need to be sure our measures of each are true to their respective constructs. A broader program of validity research should include seeking evidence of construct validity. While the relationship between ratings and value-added does examine a part of the ‘nomological net’ around the construct of teacher performance, the net could be made finer by looking at more proximate and sensitive measures of the results of practice, and relationships of ratings with other practice measures. For example, ratings on dimensions like classroom management or components like Engaging Students in Learning could be related to student survey scales measuring perceptions of classroom climate or student participation. The Gates Foundation’s Measuring Effective Teaching Project is again pioneering here by collecting student perception data that will be correlated with evaluation ratings (Bill and Melinda Gates Foundation, 2010).

It would also be useful to compare ratings from alternative evaluation systems designed to measure the same underlying practice constructs. For example, ratings on the elements of the FFT relating to content knowledge or pedagogical content knowledge could be correlated with National Board assessment certification scores that were designed to measure these constructs. A study that obtained ratings from using multiple evaluation systems on dimensions of practice each was designed to measure would provide useful construct validity evidence. Adding different methods of data collection (e.g., live observations scored based on notes versus videos) would provide the data for a multi-trait multi-method analysis that could be very convincing. There is also room for studies of the rating process. Our confidence in evaluation ratings as measures of practice would increase if we knew both that raters had followed the intended data collection, analysis, and decision-making processes and that the evidence they considered in making a rating was closely related to the practice dimension being rated.

Finally, while it would be attractive to carry out a study using a design in which teaching practice of known quality was assessed using the evaluation system under study. This is now done to some extent in evaluator training when evaluators are shown videos of teaching that have been rated by experts, and their ratings of high and low level practice is compared with the experts'. The problem is to measure "known quality" independently of the evaluation process for which we want validity evidence. While comparing ratings of experts and novices could be promising, in our research the correlation between teacher experience and evaluation ratings is not very high, suggesting that experience a poor measure of expertise.

Conclusion

Policy makers want to know both how justified they are in making inferences about teacher performance from teacher evaluation ratings, and whether the teaching practices described in a model like the FFT contribute to student learning. The evidence we have so far on the FFT is limited, but it does suggest that using evaluation scores for consequential decisions can be justified, and that at least some of the practices described by the FFT are associated with student learning, at least as the latter is represented by value-added. Clearly we need to know more about which features of a system like Cincinnati's contribute to the higher correlation between ratings and value added there, compared to our Washoe County and Coventry sites. We have speculated that this could be due to better rater training, more frequent observations, and the use of multiple raters. But there have not been any studies that vary these conditions, nor enough studies of evaluation processes that vary on them to do a meta-analysis. What we can conclude is that ratings from a system based on the FFT *can* have good reliability and agree with value-added estimates of teacher effectiveness. But of course the evaluation process has to be well designed and implemented in order to stand a chance of seeing this evidence generalize to a

new implementation by a state or school district. It seems also warranted to suggest that evaluation systems should not rely on one classroom observation, and should probably use multiple observers.

There is also much room for more and better research on evaluation systems. The Gates Measuring Effective Teaching project is likely to add massively to our small fund of validity evidence. But even after its results are published, there will still be a need to assess the validity of ratings from evaluation systems used for administrative purposes by states and school districts, to examine how evaluators make rating decisions, and to assess the benefits and costs of implementing more rigorous teacher evaluation in comparison to other methods of improving the quality of instruction.

Appendix

Inter-rater Agreement Details for Cincinnati

While the design of the evaluation process in Cincinnati precluded a direct assessment of agreement (because outside evaluators and administrators did not observe teachers on the same occasion), we estimated agreement on two domains of the Framework, the classroom environment (domain 2) and instruction (domain 3). First, we took a sample of the teachers who received a comprehensive evaluation, obtained the observation summaries completed by both types of evaluators, read the observation summaries and matching the quoted rubric language to the rubric level for each observation on each standard. We then assigned each teacher the rubric level for each standard which the preponderance of evidence supported. (For example, if the summary quoted language found at level 2 of the rubrics, we assigned a score of 2 for that standard in that observation.) Lastly, we combined standard ratings into a domain rating for each of Domains 2 and 3 by using the formula specified by the District. This produced an estimate of what the domain scores would have been had the administrator and outside evaluator had made separate, independent judgments at the domain level. Note that these estimates of agreement between evaluators are likely low because we could not separate variance due to occasion from variance due to evaluator. Using the same data, we also calculated the percentages of estimated absolute agreement on each domain. Also, because the district's method of combining scores on components to calculate a domain score resulted in an integer score, an administrator's average of 2.6 would count as a 3 while an external evaluator's average of 2.4 would count as a 2.

Inter-temporal Agreement Details for Cincinnati

Some Cincinnati teachers in our sample were rated by the same external evaluator over six observations. These were primarily first year teachers. Using the same method of

constructing ratings described above, we were able to We were able to estimate agreement across time free of a rater effect. The average reliability for across six observations for the standards of Domain 2 was .89 in 2001-02 and .82 in 2002-03. For Domain 3, the reliability of a six observation average was .84 and .76 respectively. Since the teachers observed were mostly inexperienced, these reliabilities are likely lower than what one might find for ratings of experienced teachers with more established habits of practice.

References

- Bill & Melinda Gates Foundation (2010). Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project. Available at:
http://metproject.org/downloads/Preliminary_Findings-Research_Paper.pdf
- Danielson, C. (1996). *Enhancing Professional Practice: A Framework for Teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Danielson, C. (2007). *Enhancing Professional Practice: A Framework for Teaching*. (2nd Ed.) Alexandria, VA: Association for Supervision and Curriculum Development.
- Fritsche, L., Weerasinghe, D., and Babu, A. (2003). Making the connection: Linking teacher evaluation results to district accountability systems. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April 21-25.
- 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.
- Gordon, R., Kane, T. J., & Staiger, D. O. (2006). *Identifying effective teachers using performance on the job* (Discussion Paper 2006-01). Retrieved from Brookings Institution website: http://www.brookings.edu/~media/Files/rc/papers/2006/04education_gordon/200604hamilton_1.pdf
- Dwyer C. A., & Stufflebeam, D. (1996). Teacher evaluation. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 765–786). New York, NY: MacMillan.
- Harris, D. (2011). *Value-added measures in education: What every educator needs to know*. Cambridge: Harvard University Press.
- Heneman, H.G. III, and Milanowski, A.T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education*, 17:3, 171-195.
- Jacob, B.A., and Lefgren, L. (2005). Principals as agents: Subjective performance measurement in education. National Bureau of Economic Research working Paper No. 11463. Cambridge, MA: National Bureau of Economic Research.
- Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (NBER Working Paper 15803). Retrieved from National Bureau of Economic Research website: <http://www.nber.org/papers/w15803.pdf>
- Kimball, S.M. (2002). Analysis of feedback, enabling conditions and fairness perceptions of teachers in three school districts with new standards-based evaluation systems. *Journal of Personnel Evaluation in Education* 16(4), 241-268.

- Loup, K. S., Garland, J. S., Ellett, C. D., & Rugutt, J. K. (1996). Ten years later: Findings from a replication of a study of teacher evaluation practices in our 100 largest districts. *Journal of Personnel Evaluation in Education*, 10, 203–226.
- Milanowski, A.T., and Kimball, S. M. (2005). The Relationship Between Teacher Expertise and Student Achievement: A Synthesis of Three Years of Data. Paper presented at the American Educational Research Association Annual Meeting, Montreal, Canada, April 13, 2005.
- Porter, A. C., Youngs, P., and Odden, A. (2001). Advances in teacher assessments and their uses. In V. Richardson (Ed.), Handbook of Research on Teaching, Fourth Edition. Washington, DC: American Educational Research Association, 259-297.
- Rowan, B., Correnti, R., and Miller, R.J.. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools." Teachers College Record 104, no. 8: 1525-67.
- Rogosa, D., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Education Psychology*, 76(6), 1000–1027.
- Rowan, B., Harrison, D., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal*, 105(1), 103–127.
- Sanders, W. L., & Rivers, J. C. (1996). Cumulative and residual effects of teachers on future student academic achievement. Knoxville: University of Tennessee Value-Added Research and Assessment Center.
- Sartain, L., Stoelinga, S.R., and Brown, E. (2009). Evaluation of the Excellence in Teaching Pilot year 1 report to the Joyce Foundation. Chicago, IL, Consortium on Chicago School Research, University of Chicago.
- Schacter, J., and Thum, Y.M. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, 23, 411-430.
- Schochet, P.Z., & Chiang, H.S. (2010). Error rates in measuring teacher and school performance based on student test score gains (NCEE 2010-4004). Washington, DC: National Center for Evaluation and Regional Assistance, Institute of Educational Sciences, U.S. Department of Education. Available at: <http://ncee.ed.gov>
- The New Teacher Project (2010). Teacher Evaluation 2.0. Author. Available at: <http://tntp.org/files/Teacher-Evaluation-Oct10F.pdf>
- Toch, T., Rothman, R., & Education, S. (2008). Rush to Judgment: Teacher Evaluation in Public Education. Education Sector Reports. *Education Sector*.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. Brooklyn, NY: The New Teacher Project. Retrieved from <http://widgeteffect.org/downloads/TheWidgetEffect.pdf>

Webster, W. J., Mendro, R. L., Orsak, T. H., & Weerasinghe, D. (1998, April). An application of hierarchical linear modeling to the estimation of school and teacher effects. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Wright, S. P, Horn, S.P., and Sanders, W.L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57-67.