MEASURING TEACHER PERFORMANCE
FOR STRATEGIC HUMAN CAPITAL MANAGEMENT

Anthony Milanowski
Since teachers hold key roles in education and typically comprise the largest portion of the education workforce, measuring teacher performance is one of the foundational requirements for engaging in strategic management of human capital in education. Most of the basic human capital management activities – hiring, induction, performance evaluation, and increasingly compensation – depend on measuring teacher performance; performance measures are also important for documenting teacher success. But useful measurements of teacher performance need to go beyond the paradigm of having a single administrator with minimal training making a rating of satisfactory or unsatisfactory performance based on a single observation in a teacher’s classroom.

Further, recent advances in measuring teaching practice need to be combined with the developing technology for measuring the outcomes of practice, such as value-added, because neither alone is sufficient for all human capital management needs. Outcome measures such as value-added do not provide enough information for improving teacher performance. Likewise, instructional practice measures, without linkages to impacts on learning, are unlikely to retain rigor and relevance to improving instructional effectiveness. At this stage of the policy debate, many stakeholders will accept a practice measure only if it can be linked to the “bottom line” student outcomes.

Measuring Teaching Practice

A strategy for measuring teaching practice should start by translating a vision of effective instruction into an explicit teaching performance competency model. This model would summarize the teaching behaviors and related skills that constitute effective teaching, including behaviors and skills in broad domains such as instructional planning, classroom management, delivering instruction, and what teachers need to know and be able to do to implement particular
state or district instructional strategies. (For example, a district may be emphasizing the use of instructional technology or the use of diagnostic assessment data by teacher teams.) The performance competency model then provides a template for aligning all state or district human capital management measurement and practices so that they work together to acquire, develop, motivate, and retain teachers with the requisite competencies.

Developing a teacher competency model does not require reinventing the wheel. Designers can start with state teaching standards, standards promulgated by national organizations such as the INTASC and the National Board for Professional Teaching Standards, and frameworks such as Danielson’s Framework for Teaching or the New Teacher Center’s Continuum of Teacher Development. Beginning with an existing model captures those aspects of teaching that are similar across states and districts. Designers would then add competencies that reflect a particular state or district vision and those needed to support local instructional initiatives and strategies. Several districts (e.g., Chicago and Cincinnati) and states (e.g., Idaho and Delaware) have modified the Danielson’s Framework for use in their teacher evaluation systems. Odden and Wallace (2008) presented another Framework modification, with 15 standards addressing planning, classroom management, delivering instruction, reflection on teaching, collaboration with colleagues, and communication with families.

Another approach might be to use the Classroom Assessment Scoring System (CLASS) for competencies related to classroom management, student engagement, and teacher-student interactions, and then add specific competencies related to instructional planning, student assessment, and district instructional strategies. The Measuring Effective Teaching project, supported by the Bill and Melinda Gates Foundation, is testing the use of CLASS combined with a more subject-specific assessment (e.g., in mathematics or language arts).
Measurement Systems

To assess performance relative to the competency model requires developing three measurement systems: 1) observations of classroom practice for use in periodic formal teacher evaluation, 2) teaching “work samples” or performance assessments for decisions such as granting tenure or movement on a career ladder, and 3) classroom walk-throughs that provide information for everyday performance management. All three are needed to ensure complete coverage of important competencies and to allow for different uses.

Classroom observation. To be useful for human capital management, classroom observations need to be based on a measurement system that promotes reliability and validity. Based on our reading of the literature on teacher evaluation and specifically our review of best practices in assessing teaching practice (Milanowski, Heneman, and Kimball, 2010), an observation system should include:

- Multiple competency levels defined by rating scales or rubrics providing behavioral examples of the practice levels. Rubrics guide evaluators in making more reliable decisions, provide teachers with concrete descriptions of what good performance looks like, and communicate performance expectations.

- Evidence gathering and evaluation procedures that are clearly delineated so they can be implemented uniformly.

- A focus on those aspects of instructional practice that can be observed in a typical instructional period, such as student behavior management, use of instructional time, rapport with students, student engagement, and lesson adjustment. Though observations should not ignore teacher content knowledge and application of content-specific pedagogy, these are hard to observe in a limited set of classroom observations, especially by an administrator.
who has no background in a content area. In-depth assessment of these areas should be done using teaching work samples (discussed further below).

- Observers trained in observing and recording evidence, interpreting the rubrics, and managing any biases arising from their personal views of good teaching. Training should involve the use of videos or live practice observations to help observers understand how to recognize the levels of competency and apply the rubrics. This should be followed by an assessment of observers’ ability to apply the rubrics in comparison to a set of master observers. The National Institute for Excellence in Teaching’s TAP model and Cincinnati provide this kind of evaluator training.

- Collection of artifacts such as lesson and curriculum-unit plans, assessments, and student work, which provides a basis for assessing competencies such as planning or alignment of curriculum to state standards. However, artifact collection needs to be carefully specified and limited so that teachers do not feel they need to develop a full blown portfolio.

- Multiple observations. Because teaching activities vary substantially over the day, week, and year, one observation per year is unlikely to provide a fair or representative sample of teaching practice. Although there is no magic number of observations that guarantees a good representation, three observations would seem a bare minimum. Having even two observations allows a check on how representative one observation actually is. In our own research, we found that four to five observations can give a high level of reliability.

- Multiple observers, including observers from outside the school. The teacher evaluation literature, including the recent Widget Effect study (Weisberg, Sexton, Mulhern, & Keeling, 2009), suggests that school administrator evaluations cannot be free of leniency. It may be too much to expect administrators to ignore the need to maintain harmonious relationships
and thus give marginal teachers the benefit of the doubt. Further, some administrators may never have seen really good teaching, and many at the high or middle school level may not have in-depth knowledge of all content areas. A content-knowledgeable observer from outside the school can add a more objective and informed perspective. The use of Consulting Teachers in Cincinnati and Master Teachers in Washington, D.C. are good examples of using observers from outside the school.

Many of these recommendations represent standard “best practice” as it has been developing in the performance assessment field. There is now increasing evidence that ratings from carefully designed and implemented observation systems can be sufficiently reliable and valid for consequential uses. Such evidence includes our own research on teacher evaluation systems based on the Framework for Teaching (summarized in Heneman, Milanowski, Kimball, and Odden, 2006), Kane, Taylor, Tyler and Wooten’s (2010) further study of Cincinnati’s system, and a recent study of the implementation of the Chicago Public Schools’ version of the Framework (Sartain, Stoelinga, & Brown, 2010). Reliability and validity evidence is also available for the NIET’s TAP teacher evaluation process, which draws from Danielson as well as several other assessment frameworks (Schacter & Thum, 2004, 2005; Daley & Kim, 2010). Research on the CLASS system (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008; Pianta, LaParo, & Hamre, 2008) also provides evidence for high reliability and that ratings are correlated with student outcomes.

As well as promoting reliability and validity, system design also must include features that help teachers learn from the results. Feedback should enable teachers to understand why they received the scores they did. Feedback should be specific and reference the rubric or rating scale. There should also be someone trained and available to provide coaching and assistance to
teachers who want to improve their performance. Professional development opportunities linked to the competencies should also be readily available.

The demands of improved observation systems on school administrators should also be considered. If human capital management is truly a priority, administrators’ jobs may need to be redesigned to free up time. Help from teacher leaders inside the school and evaluators from outside will be needed. The evaluation process should also be differentiated: new teachers and teachers that are struggling should receive a full dose of observation, feedback, and coaching every year, while experienced, tenured teachers might receive a full set of observations only every third year. Danielson & McGreal’s (2000) book on teacher evaluation provides a good discussion of differentiation. It is also likely that some of the video observation tools being developed by the Measuring Effective Teaching project would be big time savers, allowing an “observation” to be uploaded to a web-based system where it can be scored by trained assessors.

Performance assessments or “work samples.” This tool would be used to compliment classroom observations by providing a more in-depth assessment of content knowledge, pedagogical content knowledge, use of formative assessment data, and differentiation of instruction. Teachers would demonstrate a set of specific competencies in response to a set of prompts or questions, and by including and commenting on artifacts such as unit or lesson plans, assignments, completed student work, and assessments. Teachers could also be asked to identify successful and struggling students and describe how instruction was differentiated.

The prototypes here are the National Board for Professional Teaching Standards (NBPTS) assessments (for experienced teachers), and those developed by the Performance Assessment for California Teachers (PACT) Consortium (building on the Connecticut BEST teacher licensure system) for new teachers. These assessments can be very reliable and have also
been shown to have a positive relationship between scores and measures of student learning such as value-added (for PACT, see Wilson, Hallam, Pecheone, & Moss 2007; for the National Board, National Research Council, 2008). They can also be administered efficiently via electronically-submitted portfolios reviewed during the summer by a group of content experts trained in scoring the items. Because this approach is more intensive, it would be used mostly for major decisions like tenure or movement on a career ladder or knowledge &and skill-based pay structure.

In order to maximize efficiency, the state should probably take the lead in the developing and administering performance assessments. Not only would the state be more likely to have the resources and could use the result in multi-tier licensing systems, but also the state determines the student content standards from which teacher content knowledge should be derived.

**Classroom “walk-throughs”**. Even frequent formal observations cannot provide enough information on typical instructional practice, especially about how key instructional strategies are being routinely implemented in the classroom. Classroom walk-throughs (brief, focused visits) are more efficient for this purpose. Such walk-throughs should use a structured instrument that is aligned with the performance competency model. Walk-throughs get school leaders, instructional coaches, and mentors into the classroom frequently enough to see how teachers are developing and whether key instructional strategies are actually being implemented. If there are problems, school leaders can determine whether the problem is lack of skill (suggesting a professional development solution), lack of motivation (suggesting attention to goal setting and performance management) or context (perhaps teachers lack the time or resources). If data are collected systematically, they can be used to diagnosis and to evaluate the actions taken in response. Walk-throughs also provide opportunities to give more frequent formative feedback and encouragement to teachers and to recognize and reinforce good performance. Having school
leaders, instructional coachers, or mentors frequently in classrooms looking for major elements of an instructional strategy also sends the message that these things are important and that teachers are expected to implement them.

Many advocates of walk-throughs oppose using them to evaluate individual performance. Yet walk-throughs also have strong potential to improve summative teacher evaluation by providing a more representative sample of practice. It may be useful to designate some walk-throughs as non-evaluative (perhaps those done by instructional coaches and peers) and some as feeding into evaluations (perhaps those done by administrators).

**COMBINING TEACHING PRACTICE MEASURES WITH TEACHING PRODUCTIVITY MEASURES**

Over the past 20 years, the statistical methods that generally go by the name of “value added” have made the idea of measuring teacher productivity credible (Harris, 2011). The basic idea behind value added is to estimate the contribution of a teacher to student learning by comparing the average achievement of a teacher’s students to the level of achievement that would be expected for an average group of students with similar characteristics, including prior levels of achievement. The difference between the expected and actual level of achievement is an estimate of the “value-added” by a teacher. In practice, teacher value-added is estimated using relatively complex statistical models that predict the expected level of student achievement from prior achievement (e.g., last year’s test scores) and student characteristics such as socio-economic status, disability, and whether the student’s native language is not English. Most teacher value-added estimates are best interpreted as relative measures of productivity, since the expected achievement is generally based on the average for the jurisdiction.

While there are some major issues in using value-added that are still to be resolved, value-added methods are the best productivity measures we have. Many states are moving ahead
with them and recent Federal educational policy (notably the Race to the Top competition) has promoted the use of teaching productivity measures in human capital decision making. As districts and schools are increasingly held accountable for student achievement, some kind of outcome measure is needed that communicates the importance of getting results. At the individual teacher level, value-added is a much fairer way to measure a teacher’s contribution than the average level of student attainment. But care must be taken in how value-added estimates are used.

While we believe that using value-added estimates of classroom productivity together with assessments of teaching practice is the best foundation for making teacher human capital management decisions, we do not think that they should just be averaged into one overall measure of teacher performance. As discussed further in Odden (2011), value-added and instructional practice measures represent two different constructs and have different measurement properties. The two scores can be used together but adding them together would be like adding a person’s weight and height.

How value-added and practice measures should be combined differs with the use to be made of them. A conjoint decision rule, such as requiring a minimum score on both measures, is a natural model to use for a tenure decision or movement on a career ladder. For example, a district might require a teacher’s practice to be scored at the “proficient” level on the practice assessment by her third or fourth year, and have an average value added (based on three years of estimates) above a set level. For use in a termination decision about a tenured teacher, the analogous approach would be to identify those who have both low value-added and low teaching performance assessment scores. For both of these uses, setting cut-off points for the value-added
measure will require substantial thought, since there is no natural cut-off point that represents acceptable performance.

Although average value-added seems like an attractive cut point, there are two problems. First, in most value-added systems about half the teachers measured will be below average. This is probably too many to terminate or not tenure, simply because it would be hard to replace this many. Second, value-added estimates, like nearly all performance measures, have a degree of error. Some of those who are just below average in the value-added distribution may really above average, and some who show up as just above average may really be below. This error can be substantial (Schochet & Chiang, 2010), and it shows up as changes in teachers’ rank in the value-added distribution from year to year (Goldhaber & Hanson, 2008; McCaffrey, Sass, & Lockwood, 2008).

Thus it would be wise to use multiple years of value-added data for decisions like tenure, a pay raise or termination, and to set the lower limit at some interval below average value-added. For example, the minimum required for tenure might be set at the lower limit of the conventional 95% confidence interval. Narrower confidence intervals could be used if the decision makers were more inclined to avoid tenuring teachers who later prove to be less effective (false positives) than losing teachers who turn out to be better than predicted based on their initial performance (false negatives).

Yet another option would be to calibrate value-added in terms of the gains needed to move students to state proficiency standards. Value-added estimates could be used to develop expected trajectories for students, and value-added levels sufficient to maintain the trajectory could be required. In any case, districts would want to estimate how many teachers would be
terminated and consider whether the additional terminations could be replaced without lowering the hiring bar too far given the supply of new teachers.

Another approach for termination might be to use consistently low value-added as an initial signal that a teacher needs to be reviewed, no matter how high her or his practice assessment scores. Teachers with very low value-added – say in the bottom 15-20% of the value-added distribution for three years in a row – would have their practice reviewed by an evaluator from outside the school. If this evaluator found that practice was below the “proficient” level, the teacher would be given a year to improve practice. An improvement in practice to the “proficient” level would suffice to keep the teacher employed. Failing to improve the practice rating would lead to termination. This teacher might also be expected to have classroom value-added results above the bottom 20% for the next two years in order to avoid another outside evaluation.

Combined value-added and performance measures would strengthen the basis for giving ongoing pay raises in a knowledge and skills-based salary schedule (Odden, 2011). It would ensure that movement from one pay category to the next higher pay category would be done based on evidence of both improved classroom practice and impact on student learning gains.

For one-time pay bonuses, there is less need to directly combine value-added and teaching assessment results. The stakes are lower and so is the concern for false positives and negatives because the reward is a one-time event, and is generally not so large as to seriously affect someone’s financial situation. In this case, a natural model is a performance scorecard that simply reports the results of separate measures of performance (like a student’s report card) and associates a bonus amount with achieving performance goals on each measure. This simple approach is also easy to understand and allows teachers to be recognized for either practice or
results, or both. This approach is used to determine performance bonuses in the National Institute for Excellence in Teaching’s Teacher Advancement Program model.

**OTHER USES OF VALUE ADDED TO SUPPORT HUMAN CAPITAL MANAGEMENT PROGRAMS**

Value-added or similar productivity measures can also help improve human capital management systems, even if they are not used as direct inputs to decisions about individual teachers. Comparing the performance evaluation ratings given by evaluators with the value-added of the teachers they rate might also help to reduce the tendency toward evaluator leniency. Showing an evaluator that value-added varies a lot while most teachers receive similar evaluation ratings might help raise awareness of leniency and motivate reflection on how assessment decisions are made. Value-added teacher productivity estimates can also be used to evaluate the effectiveness of other human capital management systems. For example, evaluating the effectiveness of selection practices such as web-based screeners and interviews can be done by tracking new hires to see if high selection ratings predict high value-added. Even if value-added was not used for tenure decisions, a district could also examine its tenure criteria by seeing whether it is the more productive teachers who do receive tenure.

**SUMMARY**

This article has argued that both measures of instructional practice and teaching productivity are needed for strategic management of teacher human capital. We recommend that a state or district begin developing practice measures by translating its vision of effective instruction into an explicit teaching performance competency model, which summarizes the teaching behaviors and related skills that constitute effective teaching, and includes what teachers need to know and do to carry out state or district priority instructional strategies. The model can then become the foundation for a set of practice measures including observational
rubrics used for performance evaluation and management, performance assessments that would be part of tenure and career ladder or knowledge and skill-based pay systems, and walk-through tools used for day to day performance management and evaluation of instructional strategy implementation. These tools can then be combined with measures of teaching productivity such as value-added estimates to represent both teaching practice and its results in performance measures used for human capital management decisions. Productivity measures should also be used to evaluate the quality of teaching practice measures and the effects of human capital management programs.
References