WCER Working Paper No. 2011-2
March 2011

Teaching Assessment for Teacher Human Capital Management: Learning From the Current State of the Art

Anthony T. Milanowski
Consortium for Policy Research in Education/
Wisconsin Center for Education Research
University of Wisconsin–Madison
amilanow@.wisc.edu

Herbert G. Heneman III
Consortium for Policy Research in Education/
Wisconsin Center for Education Research
University of Wisconsin–Madison
hheneman@bus.wisc.edu

Steven M. Kimball
Consortium for Policy Research in Education/
Wisconsin Center for Education Research
University of Wisconsin–Madison
skimball@wisc.edu

WCER
Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

WCER working papers are available on the Internet at http://www.wcer.wisc.edu/publications/workingPapers/index.php. Recommended citation:

> Milanowski, A. T., Heneman, H. G., III, & Kimball, S. M. (2011). *Teaching assessment for teacher human capital management: Learning from the current state of the art* (WCER Working Paper No. 2011-2). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

The research reported in this paper was supported by the Ford Foundation and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the Ford Foundation, the institutional partners of the Consortium for Policy Research in Education, WCER, or cooperating institutions.

Teaching Assessment for Teacher Human Capital Management: Learning From the Current State of the Art

Anthony T. Milanowski, Herbert G. Heneman III, and Steven M. Kimball

Why Measure Teaching Performance?

Teacher performance in the classroom is the lifeblood of the educational enterprise. Teachers weave a combination of knowledge, skills, and abilities into specific performance competencies that become drivers of student learning and achievement. Thus, systems for assessing teaching are increasingly recognized as an important part of the instructional improvement puzzle—and specifically, of any attempt to develop a coherent system for the strategic management of teacher human capital.

The teaching competencies to be measured can be identified through a combination of research and expert judgment and then formalized into a teacher performance competency model. Usually, the competency model specifies key performance domains (e.g., delivery of classroom instruction) and, within each domain, the specific behaviors deemed desirable for effective classroom instruction (e.g., use of assessments in instruction). Examples of other common domains are instructional planning, classroom management, interactions with others (staff, parents), and professionalism.

Desired teacher behaviors, whether incorporated into a formal competency model or not, must be identified and agreed upon prior to their use in teaching assessment (such as in teacher evaluation). These competencies then become the basis for making actual assessments of teacher performance. The assessments will typically yield both numeric and qualitative information (e.g., ratings and written comments). This information, in turn, may be an input to various human resource (HR) practices within the district.

There are typically seven major teacher HR practice areas within a district:

1. Recruitment;
2. Selection;
3. Induction (preservice and after hire);
4. Mentoring;
5. Professional development;
6. Performance management; and

Revised for presentation at the American Educational Research Association annual meeting, May 2010. Comments and suggestions are welcome. Correspondence should be directed to 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 (telephone: 608-262-9872; e-mail: amilanow@.wisc.edu).

7. Compensation.

These HR practice areas can be aligned with teacher performance competencies to help a district acquire, develop, and retain a competent teacher workforce. *Vertical alignment* requires that the competencies be embedded within the HR practices (e.g., professional development activities that focus on improvement of the desired competencies). *Horizontal alignment* requires that HR practices such as teacher evaluation and professional development be mutually supportive in their competency emphasis (e.g., a teacher evaluation system that provides information about teacher competency deficiencies and is used as input for teachers' specific professional development plans).

The total teacher HR system is made up of the desired teacher performance competencies, a performance assessment process, and the HR practices themselves. Figure 1 depicts the HR system, with teacher performance competencies at the core, performance assessment serving as a linking pin between competencies and HR practice, and HR practices built on top of (vertically aligned with) the performance competencies and assessment. Because the HR practices themselves all have a common foundation, they are horizontally aligned. An aligned HR system such as this represents a truly strategic vision of HR, one that could be called the *strategic management of human capital*.

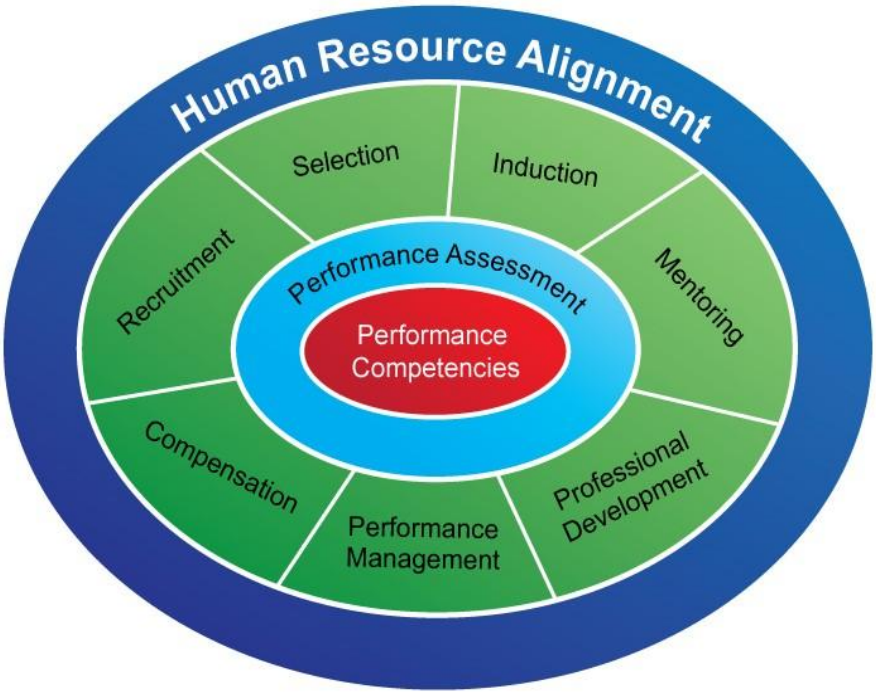

Figure 1. **HR alignment and performance assessment.**

Teaching performance assessment can be used in each of the seven HR practice areas. For example, some form of teaching assessment can be part of a district's selection process. Formative teaching assessment can be incorporated into induction and mentoring programs. Teaching assessment is obviously at the heart of performance management, which involves summative evaluation and then feedback, goal setting, and coaching for improvement where needed. The results of formative or summative teaching assessment can be used to identify

professional development needs and to evaluate the impact of professional development programs. Teachers might even be compensated based on the degree of competency development, as measured by a teaching assessment system. If based on the same competencies, assessed in a consistent manner, all of these programs will be mutually reinforcing and send the same message about what the district sees as quality instruction. When this underlying model reflects the district's vision of instruction, it provides a common language for thinking and talking about good teaching. This common language is, in turn, central to developing a culture of high instructional performance.

Some might wonder why it is necessary to measure instructional practice when value-added technology promises to measure a teacher's contribution to student achievement. While value-added estimates of teacher effectiveness are important indicators, they are not sufficient by themselves as a representation of teaching performance for use in strategic human capital management. There are several reasons for this:

1. Value-added can only be used to measure the performance of teachers of regularly tested subjects. According to one estimate (Prince et al., 2008), the student test data needed for value-added analysis is not available for about 69% of the nation's teachers.

2. Value-added estimates can usually only reliably distinguish between the highest and lowest performers—say, the top and bottom 20%. Even for these teachers, value-added estimates often differ substantially from year to year (Goldhaber & Hanson, 2008a; McCaffrey, Sass, & Lockwood, 2008).

3. Value-added estimates show only how well a given teacher's students are doing, on average, compared to other teachers' similar students. If improving value-added performance requires improving instruction, teachers and administrators need to know what specific instructional practices need to be changed.

4. Although value-added modeling can be used to evaluate, retain, and pay current teachers, it is much less clear how the approach can be used to recruit and select new teachers, identify specific professional development needs, or structure induction and mentoring programs.

Thus, we believe that value-added and behavioral measures of teaching performance will both be needed for the foreseeable future. We discuss some ways value-added modeling and teaching practice assessments can be used together in the final section of this paper and in the appendix.

Purpose of the Study

This paper reports on a study of the current state of the art in teaching assessment. The major goal of the study was to examine a sample of assessment systems and then develop a specification for a state-of the art performance assessment system to be used for human capital management functions. Our hope was that this specification would (a) provide a guidepost for working on a coherent instructional vision and devising methods to assess the alignment of actual instruction with that vision and (b) help states and districts think about how they want to develop their own teaching competency model and what assessment approaches fit best with different uses of this model. To that end, the paper concludes with an initial specification for a

high-quality, multiuse assessment system and a preliminary road map for developing such a system.

Method

To inform our thinking, we began by reviewing several prominent assessment systems. We chose eight that seemed to incorporate best practices in assessment technique, cover a range of teacher performance competencies, and represent a variety of approaches and uses. We limited our review to systems that are currently in use by states and districts and that have some evidence of effectiveness drawn from research, practice, or theory. We were interested in what the state of the art looks like and in the degree to which these systems have converged on important issues such as the key competencies for teaching and the reliable measurement of performance. We were also hoping that by choosing systems designed for a variety of purposes, we would get some insight into how a district might adapt assessment of a single underlying competency model to use for different human capital management purposes.

The eight systems selected for this study are:

- The *PRAXIS III* teacher performance assessment, developed by the Educational Testing Service for use in teacher licensure;
- The Performance Assessment for California Teachers (*PACT*), developed by a consortium of California teacher training institutions, led by Stanford University, and used for initial teacher licensure;
- The Formative Assessment System Continuum of Teacher Development (*FAS Continuum*), developed by the New Teacher Center at the University of California, Santa Cruz;
- The Framework for Teaching, as originally developed by Charlotte Danielson (*original FFT*);
- The version of the FFT adapted for teacher evaluation by the Cincinnati Public Schools (*Cincinnati TES*);
- The teacher evaluation process used by the National Institute for Excellence in Teaching's Teacher Advancement Program (*TAP*);
- The assessment system developed for certification by the National Board for Professional Teaching Standards (*NBPTS*);[1] and
- The Classroom Assessment Scoring System (*CLASS*) developed by Robert Pianta and his colleagues at the Center for Advanced Study of Teaching and Learning at the University of Virginia.

[1] The NBPTS assessments were originally developed by the Educational Testing Service. They have been further developed and are now administered by Pearson.

The CLASS, PACT, and the NBPTS assessments have different versions. We chose to examine the K–3 version of CLASS, the PACT secondary mathematics assessment, and the mathematics/early adolescence assessment of the NBPTS. We chose mathematics for both the PACT and the NBPTS assessments in order to compare different approaches to the challenge of assessing content and pedagogical content knowledge for the same subject area. We chose the K–3 version of CLASS because it is more mature than the middle/secondary version. It is also more closely related to the extensive research base of the pre-K version, but is more relevant to most districts' K–12 structure than the pre-K version.

Table 1[2] provides a brief overview of the systems we reviewed, including their original purposes, current uses, theoretical underpinnings, and basic instrumentation structures.

To begin the study, we obtained documents describing the systems, including descriptions of the performance dimensions and rubrics, training offered, and uses proposed. For the PACT and the NBPTS assessments—which use slightly different standards and rating scales for different content areas and/or student age groups—we looked at the standards and rubrics that apply to a specific subject and grade level (secondary mathematics for PACT and middle school mathematics for NBPTS). We conducted literature searches to locate any research on reliability, validity, or effects of using these systems. We contacted developers to resolve any questions we had about the systems. We then analyzed each and compared them on the following dimensions related to use in human capital management activities:

- Underlying competency model
- Assessment procedures
- Evidence for reliability and validity

Based on these comparisons, we summarized the similarities and differences. We then identified what we believed to be the best features of these systems for use as a basis for district human capital management practices.

Results

Two Basic Assessment Approaches: Performance Tasks and Observation

One major distinction between the systems we reviewed is their primary method of collecting evidence of performance competency. The NBPTS assessment and PACT are both performance assessments in the technical sense. That is, both involve defined performance tasks designed to allow teachers to demonstrate specific knowledge, skills, and abilities. It is notable that both systems also have subject- and/or grade-specific standards and rubrics. These similarities are not accidental, stemming from both purpose and design. Both of these assessments were designed for certification, and they therefore concentrate on whether teachers have the skills to perform, rather than measuring their typical performance. In addition, according to PACT developers, the NBPTS assessment process was used as a general model in

[2] Tables follow the references.

developing both PACT and its precursor, the Beginning Educator Support and Training (BEST) licensing assessment formerly used in Connecticut.

The other systems are primarily based on live observation in classrooms. Within this group, PRAXIS III, the FFT (both the original and the Cincinnati version), and TAP have a degree of family resemblance. The development of PRAXIS III influenced the development of the FFT, which in turn influenced the development of the TAP model. Charlotte Danielson, the developer of the FFT, has been involved with both the Educational Testing Service and the National Institute for Excellence in Teaching in the development of PRAXIS III and TAP, respectively. The evidence base cited by Danielson is in part drawn from the development work done for PRAXIS III, and the guidebook for TAP cites Danielson's work as one influence during the development of that system. Some concepts are found in all three systems, and similar phrases, in the rubrics. It could almost be said that the TAP system is the latest step in a development that began with PRAXIS III.

Although the FAS Continuum is not as closely related to this family of assessments, it does have some similarities with the FFT and PRAXIS III, with respect to both content and method (classroom observation). It also has a distant relationship to PACT in that it is based on the California Standards for the Teaching Profession, from which were derived the California Teaching Performance Expectations, which PACT was designed to address.

CLASS is literally in a class by itself. While based on observation, its original development was for use in research on teacher-student interactions in pre-K classroom setting, and its rigorous observation methods make it stand apart from the other approaches.

Points of Comparison

Underlying Competency Model

An important issue in developing a teaching assessment is the adequacy of the underlying competency model in reflecting both the aspects of teaching that influence student learning and the specific local strategies for improving achievement. While we believe that the competency model should reflect the local vision of instruction and local improvement strategies, there are also likely to be many competencies common across districts and states. It would seem useful for developers of an assessment system to consider assessing these. It would also seem useful for those wishing to adopt or adapt an existing system to know what competencies the system assesses. Thus, we set out to identify the competencies common to the eight systems and coverage of some specific competencies thought to be related to student learning.

We assessed the similarities and differences in underlying competency models in two ways. The first involved counting the number of performance dimensions in each system that referenced the following eight competencies often linked to improvement of student achievement:

1. Attention to student content standards;

2. Use of formative assessment to guide instruction;

3. Differentiation of instruction;

4. Fostering of student engagement;

5. Use of instructional strategies that develop higher order thinking skills;

6. Content knowledge and pedagogical content knowledge;

7. Development of personalized relationships with students; and

8. High expectations for students.

By *performance dimensions*, we mean the standards or components within the system on which teaching would normally be scored or rated. Since several systems organize aspects of teaching hierarchically, we had to decide which level to take as representing performance dimensions. For example, the original FFT has four domains, 22 components, and 66 elements but does not specify the level to be scored; we decided to treat the 22 components as the performance dimensions, thinking it unlikely that a district would want to score teaching on all 66 elements. The Cincinnati TES has 32 dimensions of teaching grouped into 15 standards; since teachers are rated only on the 15 standards, we chose these as dimensions. We chose the 10 dimensions of CLASS that are normally scored, even though the rubrics might be used to score 42 aspects of teaching. The FAS Continuum includes six standards and 32 rubric strands, and like the FFT, it does not specify the level to be scored; six standards seemed like too few, so we treated each of the 32 rubric strands as a performance dimension.

Three of the remaining systems lack a hierarchical structure: TAP has 26 rated dimensions; PRAXIS III, 19; and PACT, 12. Including the NBPTS assessments in these comparisons was complicated because the NBPTS standards differ in structure and content across certification areas, albeit with some common themes. (In contrast, PACT uses the same dimensions but words some of them differently for different subjects.) Another complication is that the NBPTS standards are not directly rated or assessed, and no rubrics define different levels of the standards. Rather, each of the four portfolio entries and six assessment center exercises to which teachers respond reflects several of the standards. Since teachers get 10 scores (four on the portfolio and one on each exercise) rather than a score on each standard, we decided to treat the four portfolio entries and six assessment center exercises as the performance dimensions in our comparison of the competency models.

Having determined which system components to treat as the performance dimensions, we then undertook an analysis of each system's coverage of the eight teaching competencies listed above. The results are shown beginning with Table 2. Each cell in Table 2 includes three numbers indicating the degree to which the system covers the teaching competency in question. In the order listed, these are:

1. The number of performance dimensions that refer to the competency in any substantial way in each set of scoring rubrics, indicating *broad coverage of the competency*;

2. The percentage of these dimensions that refer to the competency, indicating *relative emphasis on the competency*; and

3. The number of dimensions that are *predominantly* focused on the competency, indicating *likelihood that a score or rating primarily reflects the competency*.

Several conclusions can be drawn from Table 2. First, it is clear that the NBPTS assessment puts the most emphasis on content knowledge and pedagogical content knowledge. Both the portfolio entries and the assessment center exercises assess these competencies. The assessment center exercises in particular require substantial mathematics content knowledge and are essentially a test of knowledge of specific areas of mathematics. PACT refers to these competencies in three of its dimensions.

The other systems have fewer dimensions devoted to content knowledge and pedagogical content knowledge. These systems tend to assess these competencies by having observers judge whether the teacher makes content errors, identifies and emphasizes key concepts in the discipline, makes connections among concepts within the subject, and anticipates student misunderstandings. The CLASS dimensions and rubrics make minimal reference to content, reflecting its K–3 orientation and emphasis on classroom interactions. The CLASS rubrics do mention some general aspects of language development pedagogy, however, in the language development dimension.

Most of the systems covered each of the other seven competencies in at least one performance dimension. The PACT assessment places substantial emphasis on differentiation of instruction, covering this practice in rubrics for 5 of its 12 dimensions. All of the other systems reference differentiation in at least one dimension. The competency that the fewest number of systems cover is use of state or district student content standards, which are not covered by CLASS, the original FFT, the NBPTS assessment, or PRAXIS III. In contrast, the Cincinnati TES puts a lot of emphasis here, because standards have been a big part of the district's strategy for improving student achievement.

The second way we assessed the similarity of competency model content was to lay the system rubrics side by side and identify competencies that were mentioned in at least three of the systems. After reading through the rubrics, we decided which systems contained performance dimensions that were primarily based on the competency, which contained dimensions for which the competency was a major but not a dominant part of the associated rubric, and which mentioned the competency as one of several factors in a rubric. Table 3 lists these competencies, omitting the eight constructs from Table 2. It should be noted that competencies not explicitly mentioned in a set of rubrics or dimensions were not counted, even if they were arguably implicit in the rubric.[3]

Table 3 shows that there is a substantial amount of common content across the different approaches. At least five of the eight systems we reviewed have substantial content (indicated by filled and empty squares in the table) related to the following teaching competencies:

[3] In the case of CLASS, some dimensions include subdimensions that are predominantly based on some of the teaching competencies. But since the CLASS documentation describes scoring at the dimension level, the analyses reported in Tables 2 and 3 do not count a dimension as predominantly based on a competency if it is addressed by only one or two subdimensions.

- Demonstrating knowledge of students;
- Setting appropriate instructional goals;
- Communicating instructional goals;
- Demonstrating coherent instructional planning;
- Using multiple assessment methods;
- Using assessment to plan/adjust instruction;
- Building on student interests and experiences;
- Using instructional time effectively;
- Managing student behavior;
- Using effective questioning/discussion techniques;
- Providing quality feedback to students;
- Adapting the lesson/plan to the teaching situation; and
- Reflecting on practice.

Again, however, there are some striking differences. Because CLASS is based on observation alone, it has relatively little content related to planning or other out-of-classroom activities like reflection or communication with parents. Because the NBPTS assessment and PACT collect evidence using videos and written responses to prompts, they have relatively little content related to classroom management. This is logical because one video of a period selected by the teacher is not likely to provide enough evidence in this area. In general, the FFT and Cincinnati TES, TAP, and the FAS Continuum cover generic pedagogy in more detail. These systems, especially TAP and FAS, contain detailed descriptions of teaching practice. True to its origins, CLASS contains rich descriptions of teacher-student interactions, but the CLASS rubrics we reviewed also had considerable coverage of generally applicable classroom practices. PRAXIS III, because it was designed for new teacher licensure in all subjects, focuses heavily on the basics of good general pedagogy and does so with notable economy. An interesting similarity between CLASS and PACT is that both include dimensions related to teachers' efforts to develop student language (*language modeling* in CLASS; *understanding language demands and supporting academic language development* in PACT). This explicit emphasis on language development would seem a promising addition to teaching assessment, given the importance of language for all academic subjects.

Comparison of TAP and the Cincinnati TES with the FFT based on Tables 2 and 3 illustrates a potential tradeoff between comprehensiveness and focus. The FFT is more comprehensive, whereas TAP and the Cincinnati TES pare down to focus on specific aspects of

the instructional vision their authors thought important to improving student achievement, such as alignment of instruction to student academic standards in Cincinnati and higher order thinking skills in TAP. The content of the rubrics for both also contain more detailed and specific language describing teaching for conceptual understanding and differentiation. Perhaps as the price for comprehensiveness, the FFT can be schematic is some areas. The FAS Continuum is an interesting balance between comprehensiveness and specificity, with rubrics that are often more detailed than those in the original FFT. However, because of the developers' desire to represent a holistic view of teaching, some dimension descriptors and rubric language can seem redundant. Both the FAS Continuum and the original FFT put greater emphasis on student autonomy and responsibility than the other systems. The FFT uses this construct to define higher levels of practice in 9 of its 22 performance dimensions (components), whereas the FAS Continuum covers it in 5 of its 32 performance dimensions (rubric strands). CLASS also covers it in two rubric strands of one dimension.

Summarizing the content comparisons in Tables 2 and 3, we conclude that most of the systems cover many of the core competencies of teaching and most of the currently accepted drivers of student achievement. The original FFT and the FAS Continuum are the most comprehensive in coverage of teaching behavior. TAP and the Cincinnati TES are less comprehensive but arguably more focused on specific aspects of instruction likely to influence student achievement; they may represent a worthwhile tradeoff of breadth for depth by focusing on competencies key to the developers' strategies for improving student achievement. The TAP system is especially notable for the depth and specificity with which it represents desired instructional practices. The NBPTS and PACT assessments, though emphasizing content and pedagogical content knowledge, also do a reasonable job of covering more generic core aspects of teaching, with the exception of classroom management. PACT stands out for simultaneously customizing by subject matter and retaining similar performance dimensions across assessment areas.

Assessment Procedures

Table 4 summarizes the basic features of the assessment processes used or recommended for each approach. When considering the similarities and differences among assessment procedures, it will be helpful to bear in mind some important differences between the two approaches that use performance tasks (NBPTS and PACT) and the approaches based largely on passive observation and artifact collection. Some of these differences stem from the purposes of the systems, but others are related to the assessment technology employed. Because the performance task–based NBPTS and PACT systems present teachers with a standardized set of tasks or prompts, they can be structured to ensure that teachers have to demonstrate specific skills. In contrast, systems based on classroom observation must typically take whatever evidence is observed (or is shown by the artifacts collected). Some skills may simply not be observed because the teacher does not have to use them during the observation. The potential weakness of standardized tasks is that teachers have the opportunity to prepare their responses in advance, and thus what assessors see may represent peak rather than typical performance.[4] Given

[4] This may be less an issue in practice with PACT than NBPTS. According to PACT staff, some teacher preparation programs limit the time provided to prepare the teaching event, and thus teachers in these programs may not have much time to write and polish their commentaries.

the purposes of the NBPTS and PACT assessments, this is appropriate. Observation-based systems are more likely to capture typical performance. However, even when supplemented by artifacts like lesson plans and student work, they may miss evidence relevant to some performance dimensions unless many observations are used. This limitation needs to be taken into account when assessments based on passive observation are used for consequential decisions.

The NBPTS and PACT emphasis on content and pedagogical content knowledge—noted earlier in relation to competency models—is also reflected in these systems' assessment procedures. The NBPTS uses more than 30 different assessments and different rubric language for different content areas. PACT customizes the rubric language for three of its dimensions, while using similar performance tasks across areas. The other five systems use the same basic observation methods and rubrics across all content areas.

Despite the major divide between the NBPTS and PACT assessments and the others, there are several commonalities among the systems. First, and perhaps most obvious, all have a set of specific performance dimensions or standards. While the documentation for most of the systems explicitly recognizes that teaching is a complex activity involving multiple intertwined competencies, all do break teaching down into more or less separable dimensions on which teaching practice can be scored. Compared to a single holistic judgment of competence or performance, this allows strengths and weaknesses to be distinguished and recognized or remediated. It also makes the assessors' job substantially easier because the required judgments are broken down and focused on specific types of evidence, thereby promoting reliable scoring. A system requiring a global rating of performance would allow each assessor to weigh the various aspects of teaching differently For example, one assessor might believe that no teacher can be competent without nearly perfect classroom management, whereas another might believe that a high level of student engagement can make up for imperfect classroom management.

Although all of the systems we reviewed have multiple performance dimensions or standards, they differ in the number of assessment dimensions. This is related to the breadth of the competencies covered and the "granularity" at which teaching is analyzed. All other things being equal, the greater the number of dimensions, the greater the breadth and the more fine-grained the analysis. A finer grained system provides teachers with more information on desired performance and more accurately reflects their strengths and weaknesses. But this comes at the expense of more complex scoring. To recap, the original FFT has the largest number of dimensions, with a total of 66 elements that can be rated, reflecting the comprehensiveness of the system's competency model, which includes several types of competency not found in the other systems (e.g., supervision of paraprofessionals and volunteers). The FFT also analyzes the generic process of teaching at a relatively fine level, at the expense of including many potential dimensions and perhaps diluting the message of what is important. The Cincinnati TES concentrates on 15 dimensions for rating, though the rubric structure distinguishes 32 aspects of teaching. The FAS Continuum can also distinguish 32 dimensions, but some seem to overlap or refer to multiple constructs. TAP has 26 dimensions on which ratings are given, but most contain multiple constructs. CLASS has 10 dimensions that are rated, but multiple strands within the rubrics distinguish 42 aspects of teaching. CLASS uses many of these strands to assess social interactions within the classroom, rather than try for breadth of coverage. PRAXIS III has 19 ratable dimensions; PACT has 12; and the NBPTS assessment, 10. These assessments have a

narrower focus on specific career stages (novice teachers, accomplished teachers) and emphasize the content needed to certify.

The second commonality in the systems is that they all have multilevel rubrics or rating scales with more or less specific examples of behaviors that help to define the levels. Rubrics serve a number of important functions. They are important for assessment reliability and validity; they provide guidance to assessors on what specific behaviors constitute evidence for performance at each level; and they help teachers by communicating the specifics of the vision of instruction the district wants implemented and providing concrete examples of what it would take to improve ratings. Multiple levels provide for growth compared with a simple satisfactory/unsatisfactory distinction.

It should be noted that the NBPTS rubrics are not as behaviorally based as the other systems. In particular, the scoring guide describes the performance levels in terms of the characteristics of the evidence rather than specific teacher behavior. A Level 4 response is defined as one that provides clear, consistent, and convincing evidence that the teacher has a set of competencies; a Level 3 response provides clear evidence; Level 2, limited evidence; and Level 1, no evidence. Consistent with the purpose of the assessment as an indicator of accomplished teaching, the scoring system does not define lower levels of teaching, but only the degree of evidence that teaching is accomplished. Thus, the system does not describe a developmental sequence of teaching competency like the other rubrics.

A third important similarity is that almost all the approaches pay close attention to assessor reliability by requiring assessor training and accountability. All of the systems recommend substantial multiday training in the use of the assessment tools, including the rubrics. Four of the systems (the Cincinnati TES, CLASS, PRAXIS III, and TAP) require novice assessors to demonstrate performance consistent with that of master assessors before they are allowed to serve as assessors themselves. The National Institute for Excellence in Teaching even requires users of the TAP system to provide refresher training and calibration. Two systems (NBPTS and PACT) have a sample of assessments re-rated by a second assessor to allow examination of assessor calibration. Assessors who deviate from the standards are retrained until they match a criterion or are dismissed.

Some of the approaches have additional features intended to promote reliability and validity. For example, the NBPTS assessments require that assessors have experience teaching in the subject and grade level of the assessment they score. PACT requires subject expertise, which could include experience teaching the content or training teachers in the content area. Cincinnati has also tried to match the expertise of its assessors with the subject and grade level of the teacher being assessed. The district also uses assessors from outside the school—peer teachers released from teaching for a period to specialize in teaching assessment—to evaluate first-year teachers, struggling teachers, experienced teachers in their third year, and other teachers for whom consequential decisions will be made based on results. For some uses, Cincinnati brings in multiple assessors, typically one of these specialists and a school administrator. The TAP system also uses multiple assessors, but these are usually from the same school as the teacher being evaluated. Assessors include both a master and mentor teacher and a school administrator. Mentor teachers are likely to have experience in the grade level or subject of the teacher being assessed. CLASS also recommends the use of multiple assessors, even though the system was

not designed to be used for consequential decisions. The use of outside or multiple assessors is a promising approach to reducing leniency and improving validity in summative assessments. Often, school administrators do not have the time, subject matter expertise, or motivation to do a thorough assessment by themselves. Sharing the burden of assessment may also make it easier to collect comprehensive information and make tough calls.

Some of the observation-based assessment approaches require multiple observations, another feature that should improve reliability and validity. Given that both research (e.g., Rowan, Harrison, & Hayes, 2004; Rogosa, Floden, & Willett, 1984) and experience suggest that teaching is highly variable over time, it is unlikely that one observation can provide a representative basis for an assessment, especially for consequences. TAP requires that four to six observations be made each year, with at least half of these unannounced. The Cincinnati TES has required from four to six observations for a comprehensive evaluation (e.g., for consequential decisions), with three of these unannounced. CLASS does not specify the number of separate occasions of observation, but the procedural handbook does recommend a minimum of a 2-hour observation session, in which observation and coding cycles alternate (20 minutes of observation, then 10 minutes of coding). The FAS Continuum recommends three classroom visits spread over several days or a week. Formal observation at least twice per year is advised, with shorter informal observations monthly. PRAXIS III could be used with only one observation.

There are some other differences in assessment processes, most of which stem from the intended use or from the performance task/observation distinction. The two assessments using performance tasks, PACT and the NBPTS, collect data using videos. The videos are not merely substitutes for live classroom observations. They are intended to show specific aspects of teaching rather than a random sample of teacher behavior. Moreover, the videos are not the main source of evidence in either system. PACT and NBPTS also rely heavily on what teachers write about their practice, guided by prompts designed to elicit content knowledge and content-related pedagogy; the videos serve to illustrate and confirm what teachers write. The NBPTS assessment goes even farther by including six written exercises that are essentially tests of content and pedagogical content knowledge. This approach is much more efficient than making several live observations in the hope of seeing specific skills displayed. In the course of even several observations of conventional length (say one class period), the assessor is not likely to see a full range of content or pedagogical content knowledge demonstrated.

Although all the other systems except CLASS reference content errors and content-appropriate pedagogy in their rubrics, it is hard to assess the depth of a teacher's content knowledge by simply observing. It is likely that collection and analysis of artifacts such as unit and lesson plans, student assignments, and tests would also be needed. If one also asked teachers to comment on or relate these artifacts to performance dimensions or content-based instructional goals, one would obtain even more useful evidence. But this would be moving toward the performance task method of data collection used in PACT and the NBPTS assessments. At the least, observing and interpreting this evidence to make a valid assessment would seem to require that assessors be knowledgeable about the content area. As noted earlier, the Cincinnati TES attempts to match assessors with teachers based on grade and content specialty, and the mentor teachers who serve as one set of evaluators for TAP are likely to have similar grade-level or subject experience, but the other observation-based systems do not appear to address this.

Evidence for Reliability and Validity

For use in making consequential human capital management decisions, assessment results must be reliable and valid. In this section, we look at the evidence for the reliability and validity of assessments made using each of the systems we reviewed and then identify the characteristics of the systems with higher levels of reliability and validity evidence.

Reliability. Reliability is often considered a basic requirement for teaching performance assessments used for consequential decisions. The aspect of reliability typically of most concern with these assessments is interrater reliability, often represented by interrater agreement. There is evidence on interrater agreement for five of the assessments: CLASS (Pianta, LaParo, & Hamre, 2008), Cincinnati TES (Milanowski & White, 2001; Heneman & Milanowski, 2003), NBPTS (National Research Council [NRC], 2008), PACT (Pecheone & Chung, 2007), and TAP (Schacter & Thum, 2004, 2005). Interrater agreement for these assessments is generally adequate to excellent, with PACT, NBPTS, and CLASS having the highest levels. Review of the research on these systems suggests that assessment systems can be designed to have reliability sufficient for use in making human capital management decisions with important consequences for teachers.

There are notable commonalities across these systems that we believe contribute to assessment score reliability. First, all use trained assessors, and all have some way to hold assessors accountable for applying rubrics correctly. Second, the rubrics have been developed to be specific, with well-defined distinctions among levels. Several additional factors help explain the especially high reliability of CLASS, PACT, and NBPTS scores. The PACT and NBPTS assessments involve review of standardized artifacts, and teachers are given considerable guidance on what to prepare and submit. Assessors also do their scoring away from the classroom and its distractions; thus, there is less "noise" in what is assessed than might be typical in a classroom observation. With regard to CLASS, the standardization of observation procedures likely contributes to high reliability, along with the substantial amount of rater training and the careful design of the rubrics. In addition, the CLASS reliability studies (e.g., Pianta, LaParo, & Hamre, 2008) used researchers as raters rather than school administrators, as did the TAP reliability studies. The Cincinnati TES has likewise used assessors from outside the school. These considerations suggest that high reliability is most likely when rubrics are carefully designed, procedures standardized, and well-trained outside assessors used.

Validity. Assessment validity is a more complex matter to judge. There are a number of types of evidence of validity, and two have been the focus of research on these assessment systems. One is *content-related validity* evidence, which in this context refers to the extent to which the content of the assessment (what it measures) matches the content of the teaching role. This evidence typically involves expert judgments on how well the assessment method covers or represents what teachers actually do. Extensive content validity evidence exists for PRAXIS III (Dwyer, 1994; Reynolds, 1995). This evidence arguably also applies to the original FFT (Danielson, 1996, 2007), which was built around some of the same competencies. There is no independent content validity evidence for the related Cincinnati TES, though extensive involvement of teachers and the teacher association suggests that it represents an agreed-upon vision of instruction. Substantial content validity evidence also exists for the NBPTS assessments, but in this case the evidence showed the extent to which the content of the

assessment matched the content of the NBPTS standards for teaching rather than the content of the teaching job or role as such (see Jaeger, 1998; Moss, 2008). Content validity studies were also done for PACT, showing that this assessment represents both performance dimensions important to teaching and alignment with the California Teaching Performance Expectations (Pecheone & Chung, 2007). The FAS Continuum was derived from the California teaching standards, which were themselves the subject of a content validation study (Whittaker, Synder, & Freeman, 2001). However, there does not appear to have been a formal study linking the Continuum to these standards, likely because the intended use is formative rather than summative. Content validity studies linking CLASS to state standards are under way. No content validity evidence was located pertaining to TAP.

Review of this evidence suggests that PRAXIS III and arguably the original FFT adequately reflect the most important aspects of the teaching role, though in its attempt to be comprehensive, the FFT adds some aspects that may not be important in all contexts. Note that the content validity evidence for PACT and the NBPTS is relevant to these assessments' coverage of teaching standards: idealized conceptions of what teaching should be. The applicability of their content validity evidence depends on the degree to which the state or district wanting to use these assessments believes that these teaching standards reflect the kind of teaching they want to encourage. For an assessment system to be used in a strategic human capital management system, a district or state would likely want to do its own content validity study to ensure that the standards and rubrics match its vision of teaching and the instructional strategies for improving student achievement it has chosen. Content might need to be added, as has occurred with the Cincinnati TES. The analyses described in the Underlying Competency Model section should be of use in starting to think about how a state or district assessment might capture the instructional vision.

A second type of validity evidence is the relationship between teachers' scores or ratings on an assessment and the average achievement of the students they teach. This is sometimes called *criterion-related validity* evidence. The idea is that there may be an external standard of performance (the criterion) that the assessment scores should correlate with or predict. For measures of teaching, the currently accepted criterion is the average value added to student achievement in the classrooms in which the teaching is being measured.

Criterion-related validity evidence is available for four of the assessments we reviewed: CLASS, the original FFT, the Cincinnati TES, NBPTS, and TAP. The most research has been done on the NBPTS assessments. According to a summary of the research by the NRC (2008), the average value added was consistently higher in the classrooms of NBPTS-certified teachers than in those of applicants who failed the assessment. Comparing certified teachers to non-applicants, the NRC (2008) again found the results to be mostly in favor of certified teachers, but not as consistently. With regard to the FFT, members of our research group have obtained evidence of a positive correlation between rating and classroom value added in the Washoe County (NV) school district (Kimball, White, Milanowski, & Borman, 2004; Milanowski, Kimball, & Odden, 2005). We also found stronger relationships for the Cincinnati TES (Milanowski, 2004; Milanowski et al., 2005) and a version used in a charter school Gallagher, 2004). Later research by Kane, Taylor, Tyler, and Wooten (2010) also found a positive and significant relationship between performance ratings and student achievement in Cincinnati. With respect to TAP, research by Schacter and Thum (2004) on the TAP evaluation model found

that the teachers' ratings on the TAP observation rubrics were associated with higher classroom value added, but the sample size of this study was small, and the raters were researchers rather than the school administrators or teacher leaders who normally evaluate in TAP schools. However, Daley and Kim (2010) also found a positive relationship between observational ratings of practice using the TAP model and classroom value added in a larger sample of 449 teachers. The fairly substantial amount of research on the pre-K version of CLASS (see Pianta, LaParo, & Hamre, 2008) has found consistent positive associations between the assessment's measures and student outcomes. Initial research with older students has shown some positive associations between emotional quality and instructional quality scores and reading and math achievement (Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008). Further research using the K–3 and 5–12 versions is just getting under way. Research has also been conducted on the precursor to the PACT assessment, the Connecticut BEST licensing assessment. This research found that teachers scoring higher on the BEST assessment had higher classroom value added in reading (Wilson, Hallam, Pecheone, & Moss, 2007). Similar research is currently being conducted on PACT. We found no research of this type on PRAXIS III or the FAS Continuum.

Our reading of the research is that teaching assessment systems can produce scores with useable levels of criterion-related validity.[5] Although the research is still too thin to draw definitive conclusions about what system characteristics are associated with higher score validity, we do know based on measurement theory that high reliability and substantial score differentiation among teachers are needed. It is notable that the CLASS, Cincinnati TES, NBPTS, PACT, and TAP approaches all have features aimed at promoting reliability, and all have substantial content relating to important instructional influences on student achievement. All except TAP use raters from outside the school, which may reduce leniency and contribute to score differentiation. There is reason to believe that having school administrators as the primary assessors limits score differentiation. A recent report by the New Teacher Project (Weisberg, Sexton, Mulhern, & Keeling, 2009) is the latest documentation of the well-known tendency for performance evaluation ratings to fail to differentiate much among teachers, with few rated in the lower categories (see also Dwyer & Stufflebeam, 1996; Loup, Garland, Ellett, & Rugutt, 1996; Kimball & Milanowski, 2009.) Leniency of ratings done by supervisors is also well known in the private sector (Murphy & Cleveland, 1995; Levy & Williams, 2004). Not only does leniency tend to lower correlations between teacher assessment scores and student outcome measures, but it also defeats many of the human capital management purposes for assessment scores. The pervasiveness of leniency suggests that an assessment system intended to produce ratings to be used as inputs for consequential decisions cannot depend solely on school administrators as assessors.

[5] It should be noted that most of the relationships between assessment scores and student achievement are small to moderate in strength. It is unlikely that a very strong relationship between teaching assessment scores and average value-added student achievement would be found even for the best assessment system. Not only is there measurement error in both teaching assessment scores and student test scores, but variations in student motivation, imperfect alignment of the curriculum with the tests, and misalignment between the pre- and posttests used in the value-added analyses all tend to attenuate any positive relationship. Also, strictly speaking, all reliability and validity evidence pertains to specific scores, not to the systems themselves. One reason for this is that the scores typically depend on how the assessment system was implemented. If assessors are poorly trained, do not observe carefully, or do not follow scoring directions, reliability and validity results are not likely to match those found in the research. To achieve comparable score reliability or validity, the process must be followed as designed. 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.

Toward a Specification for Teaching Assessment Tools

Based on our review of the competencies covered by the eight systems, their assessment procedures, and the evidence for the reliability and validity of assessment scores, we propose in this section a preliminary set of specifications that states or districts may want to consider when deciding what approach to take to their own assessment of teaching. Our intent is to present what a state-of-the-art teaching assessment system would look like, based on the best features of the systems we reviewed. However, we need to emphasize a few cautions.

First, no matter how well-designed the competency model and the assessment processes, no assessment system will realize the desired benefits unless implemented as intended. Moreover, it is clear from the research on performance evaluation that implementation poses a challenge for organizations in both education (Halverson, Kelley, & Kimball, 2004; Heneman & Milanowski, 2003; Davis, Pool, & Mits-Cash, 2000) and other sectors (Bretz, Milkovich, & Read, 1992; Roberts, 1995). Therefore, districts need to specify the details of implementation and develop an implementation plan, including identifying a champion for the system. They will also need to provide the various actors in the system with the resources to implement as intended and hold them accountable for doing so.

Second, there is unlikely to be one best data collection approach for all of the human capital management uses of teaching assessment. High-stakes uses (e.g., career ladders or knowledge- and skill-based pay systems) require standardization, which is more easily accomplished by using performance tasks and a fixed group of trained assessors from outside the school. Performance management and some developmental uses require "real-time" assessment of typical performance, for which observation by a local supervisor or mentor is needed. When the goal is to assess depth of content knowledge (i.e., more than the absence of content errors), asking teachers to demonstrate specific knowledge on a performance task is likely to be more efficient than making multiple observations. To effectively monitor implementation of the instructional strategy, recognize good work, and keep the focus on important performance goals, many short, unannounced observations ("walk-throughs") might be more useful than a few full-period formal observations.

Third, although the use of multiple data collection methods tailored to specific purposes is desirable, districts need to make sure that all of the methods are based on a single competency model in order to preserve the alignment of the system. Our experience in researching teacher and principal evaluation suggests that failure to do so communicates that the district at best is not serious about any one set of competencies and at worst is confused about what it is doing.

With these cautions in mind, we outline below eight specifications for a state-of-the art district teaching assessment system. As our work progresses, we intend to add more specifications related to administrative feasibility and teacher/administrator acceptance.

1. The system should be based on a competency model that includes the drivers of student achievement and the things teachers need to know and be able to do to effectuate the district's strategies for improving student achievement.

2. The competency model and the basic concepts in the rubrics need to be applicable to all grade levels, career levels, and subjects, yet customizable, as needed, to specify grade- or subject-specific instructional strategies or skills. The PACT assessments are a good example of how a single set of basic performance dimensions (in PACT, "guiding questions") can be customized to apply to different subject and grade levels.

3. If the intention is to assess teacher content knowledge and pedagogical content knowledge for consequential decisions, the assessment system should include standardized performance tasks that ask teachers to demonstrate this knowledge. Obtaining a good sample of teacher behaviors indicating this depth of knowledge requires too many observations. Moreover, even if several observations are conducted, there is no guarantee that they will yield a representative range of content-related knowledge. A standardized set of performance tasks can be planned to assess the same key content-related knowledge for all appropriate teachers.

4. Content-knowledgeable assessors with experience in the relevant grade levels should be used to judge teaching performance. The use of such assessors not only promotes more valid judgments, but also adds credibility to both ratings and any advice or coaching provided using the assessment.

5. When assessment results are used for consequential or summative purposes, the system should include features that promote reliable and valid measurement. These include (a) multilevel, behaviorally anchored rating scales or rubrics; (b) assessor training; (c) a process for determining assessor skills (e.g., requiring novice assessors to make ratings that agree with those of an expert panel before they are allowed to assess); (d) a way to hold assessors accountable for following the process; (e) use of multiple assessors or checks of assessment scoring by additional assessors, at least for a sample of decisions; and (f) use of multiple observations if assessment is primarily based on classroom observation and the intent is to measure typical performance. Assessment systems intended to be used for consequential decisions such as tenure or career ladder progression should not depend solely on teachers' direct supervisors (e.g., the principal) as assessors, but rather involve ratings from assessors from outside the school.

6. The system should include features that promote teacher learning. Specifically, assessment results should include enough specific detail for teachers to understand why they received the scores they did; someone should be trained and responsible for providing coaching and assistance to teachers who want to improve their assessed performance; and if the assessment is to be used in induction and intensive professional development, it should be embedded in a planned set of developmental activities.

7. The assessment process should be standardized and documented. In particular, the procedures for gathering, interpreting, and evaluating evidence should be spelled out so that they can be implemented uniformly, and assessment conditions should be specified (e.g., number and length of observations, length of video clips, prompts for performance tasks).

8. The system should make use of technology (e.g., web-based data collection and scoring; video) to minimize workload and improve administrative feasibility. Technology-enabled mentoring such as that available to users of CLASS (Hadden & Pianta, 2006) should be

included to help provide high-quality professional development directly related to the competencies being assessed.

An Approach to Designing a Teaching Assessment for Human Capital Management

Developing a teaching assessment for use in a strategic human capital management system is a complex undertaking. Below, we outline a process that we believe can help districts and states think about how to get started and what resources are needed. We are presupposing that the eventual goal is to use the assessment system as the basis for multiple human capital management decisions ranging from teacher selection to compensation. Note that by *assessment system*, we mean not just a single assessment, but a group of related assessments based on a single competency model that would be used for various human capital management purposes.

The approach we propose involves the following seven steps:

1. Develop a competency model on which to base the assessment system and other human capital management programs.
2. Decide on a high-leverage use of the assessment for the initial development effort.
3. Develop the assessment for the initial use.
4. Pilot-test and revise the system.
5. Analyze other human capital management assessment needs and develop supporting assessments.
6. Collect reliability and validity evidence from post-pilot administration.
7. Consider using both teaching practice assessment and measures of student outcomes (e.g., value-added estimates of classroom productivity) for human capital management decisions.

Below, we discuss each of these steps in detail.

Step 1: Develop Competency Model as Basis for Assessment System and Other Human Capital Management Programs

The design process should begin with a review of the district or state vision of instruction and strategies for improving student achievement. A group of people knowledgeable about the vision and strategies—and their implementation—can be convened to develop a list of what teachers need to know and be able to do to carry out the vision and strategies. This group should also review existing competency models, including those underlying the assessments we have reviewed; the current state teaching standards; and the district's teacher evaluation system to see how well they capture the competencies needed to carry out the vision and strategies. Using these resources, the group can develop a competency model that reflects the most important aspects of the vision and can drive successful implementation of the strategies. This model should be concise, focusing on only the most critical competencies. The model would then be shared with appropriate stakeholders for review and comment, and needed modifications made.

It is likely to be most efficient to adapt an existing competency model. Although most existing systems are likely to require some modifications to fit individual state or district needs and specific performance improvement strategies, a good deal of basic content will likely be applicable. Developing a statewide or multistate model and then customizing it to the needs of individual districts seems a plausible and economical approach. A system to assess core performance competencies could be drawn up for use by multiple districts, with non-core content or supplementary assessment procedures added for local use. The PACT consortium has taken this approach. The PACT licensure assessment is aimed at a core of the California Teaching Standards, while a complementary set of Embedded Signature Assessments are under development to address competencies specific to the teacher preparation programs of the consortium's membership. The version of the FFT used by Cincinnati is another example. Cincinnati made fairly substantial changes to the original FFT to reduce the number of performance dimensions and prioritize the teaching practices considered most important to improving achievement. Although Cincinnati might have been able to develop its system from scratch, beginning with the FFT jumpstarted the development process and avoided reinventing the wheel.

Step 2: Decide on High-Leverage Initial Use

The second step involves selecting a high-leverage use of the assessment as part of the initial development effort. By *high-leverage use*, we mean a use tied to decisions about important human capital management matters such as teacher selection, professional licensure, tenure, or career ladder movement.

Step 3: Develop Assessment Around Initial Use

Step 3 involves developing an assessment plan to specify the uses for the assessment, the competencies to be assessed, and the methods of collecting evidence on and rating the competencies. Different competencies are best assessed using different forms of evidence collection, whether performance tasks, videos, artifacts such as lesson plans and student work, or live observations. Selection of methods should take into account their efficiency and their potential reliability and validity under operational conditions.

As an example, assume a district wants to develop an assessment to be used in tenure decisions. A critical requirement for such a high-leverage use is high reliability and validity. This suggests that the assessment should be externally scored by trained assessors with subject- and grade-level expertise if possible. An assessment based on a set of performance tasks such as the NBPTS or PACT assessments would be a good candidate for measuring major competencies. One approach would be to focus the data collection on an instructional unit. This would provide a true work sample and would allow evidence to be collected about most of the key teaching competencies. In this sort of assessment, the teacher might be asked to describe the unit's goals; relate them to state or district standards; provide a unit plan and a few sample lesson plans; submit sample materials, assignments, and assessments; and describe how instruction would be differentiated for high-performing, average, and struggling students. A video could be included showing how the teacher introduced the unit and taught a key concept. Teachers could also be asked to explain their decisions and reflect on the success of the unit. To ensure that the work sample is representative, teachers could be asked to submit this material on more than one unit.

Since the assessment method just described could not get at all of the aspects of instruction that are important for a tenure decision, it would be advisable to add additional methods. In particular, one would like to see more evidence on classroom management and teacher relationships with students than would be provided by a few relatively short, edited videos. Classroom observations might be the best method to obtain this evidence, using an observational rubric such as CLASS or an adaptation of the relevant parts of the FFT.

Step 4: Pilot-Test and Revise

Assessment systems are likely to have a number of glitches and implementation problems that will show up only when they are used. It is therefore important to plan for a pilot test of the system under as close to operational conditions as possible, but without consequences to teachers. This pilot study could examine the fidelity of implementation, the feasibility of administration, the workload imposed on administrators and participants, measurement properties such as reliability and discrimination, and teacher, administrator, and assessor reactions. This information would surface most implementation problems. Design changes could then be made before the assessment is used for consequential decisions.

Step 5: Analyze Assessment Needs and Develop Supporting Assessments

We strongly recommend that the assessment system be incorporated into, and aligned with, the district's human capital management system. This will leverage the investment made in assessment and help to achieve the benefits of alignment discussed at the beginning of this paper. Note, however, that one assessment method may not be appropriate for all human capital management uses. For example, performance tasks such as those used by PACT and NBPTS do not measure typical behavior, so such assessments are not likely to be useful in monitoring everyday teaching performance.

Using the competency model, system developers need to consider what information is needed for other human capital management uses and choose a data collection system that best fits those uses. Table 5 presents some suggestions. Note that the key to developing an aligned system is to work from the competency model in designing each assessment. It is not essential that all of the assessments focus on all of the competencies; some might focus on only a subset. What *is* important is the alignment, and the advantages of alignment are lost if a district uses a teacher evaluation system based on one set of competencies, a walk-through protocol based on another, and an induction process based on yet another. When different systems are developed independently by different central office departments based on different visions of instruction, the result can be duplication of effort and confusion among teachers and school leaders about what the district values.

It may be useful for a district to consider analyzing the alignment of the entire human capital management system as part of the process of aligning assessments. Doing so would help identify all uses of teaching assessment and also surface potential misalignments, such as recruitment programs that do not supply job applicants with the basic competencies or professional development programs that do not support teachers in acquiring competencies needed for tenure or career progression. The HR alignment analysis process developed by Heneman and Milanowski (2004, 2009) may be a useful model for a district alignment analysis.

Step 6: Collect Reliability and Validity Evidence

During the first full administration of the assessment and periodically afterward, system designers should collect information on reliability of scores, such as interrater agreement. Reliability is important not only for achieving valid measurement, but also for establishing the credibility of the system with teachers. Prior experiences with teaching practice assessment, such as performance evaluations, may have convinced many teachers that practice ratings depend as much on the observer as the observed. Collecting and reporting reliability evidence may help counteract teacher skepticism about whether practice can be fairly judged.

The validity of scores should also be studied, especially the relationship between scores and other measures of performance such as value added. There should be a positive relationship between teaching assessment scores and value-added estimates of classroom productivity. Some substantive positive relationship should be found if the assessment system is focusing on the right competencies and is being used in a reliable and accurate way. We would expect to see correlations between assessment scores and value-added estimates in the .2–.5 range. Correlations of this size are meaningful in terms of the long-run improvement of faculties if assessment scores are used for human capital management decisions such as tenure, compensation, and remediation. Smaller correlations are evidence that either the assessment system is not focusing on the most important drivers of student achievement or the measurement procedures are not reliable or being implemented as intended. Calculating these correlations provides districts with important evidence that can be marshaled to justify the use of teaching assessments for human capital management decisions and to improve the assessment system. For example, it may be found that the classrooms of teachers with low assessment scores have low value added, but that among teachers with high scores, value added varies substantially. This could be evidence of assessor leniency or failure of the assessment system to adequately represent the teaching practices that contribute to large learning gains. In this situation, it would be advisable to investigate. Assessors could be interviewed and assessment scores rechecked (especially if assessments were made using videos and artifacts) to look for leniency. High-rated teachers could be interviewed or observed to look for practices that differentiate high and low value-added classrooms but are not be reflected in the assessment system.

Step 7: Consider Using Both Teaching Practice and Value-Added Measures

At the beginning of this paper, we argued that measuring teaching using only estimates of value-added student achievement was neither sufficient to improve instruction nor sufficiently valid for all human capital management uses. However, outcome measures such as value added have important roles to play in teaching assessment. Clearly, student achievement is currently the paramount objective of districts and schools. Since districts and schools are being evaluated based on student achievement results, many want to communicate the importance of achievement to teachers by evaluating their individual contribution as well. Further, if student achievement is the paramount goal, it seems a logical step to assess teachers' contributions to student achievement and make at least some human capital management decisions based on that contribution. Considering also the persistent problem of leniency in evaluations of practice based on administrator judgment (see Weisberg et al., 2009, for the latest documentation), it is also clear that such evaluations are typically insufficient for human capital management uses such as paying for performance or granting tenure.

Two approaches to using value-added and similar outcome measures are calibration and complementary measurement. *Calibration* refers to the use of value-added estimates to examine and improve the judgments of individual assessors or groups of assessors. The basic idea is to analyze the relationship between assessment scores and value-added estimates for single assessors or assessor groups in order to identify differences that could indicate assessor problems in applying the process. For example, an assessor may rate most teachers at the highest level, even though the teachers have widely varying value added. This would suggest that the assessor may not be using the high end of the rating scales correctly (assuming that the rating scales do a good job of representing teaching practices that are associated with student achievement). Such an assessor could be interviewed to see how well he or she was applying the system, and if problems become apparent, additional training could be provided. Such interviews can provide clues about why assessors may not be following the process (Kimball & Milanowski, 2009). Even simply showing assessors graphs of the relationship between value-added estimates and the teaching practice scores they assigned could help raise awareness of leniency and motivate reflection on how assessment decisions are made. Under certain conditions, it also might be possible to adjust the ratings of particularly lenient or severe assessors.

Complementary measurement refers to using both value-added and judgment-based practice measures for some important human capital management decisions. Using both provides multiple indicators of teaching performance, recognizing the importance of both outcomes and teaching practice. For example, the granting of tenure or movement to the next level of a career ladder could require both the attainment of a certain score on the teaching performance assessment and a consistent pattern of value added, perhaps 3 years of positive value-added estimates (which would show the teacher's classroom was consistently achieving above-average learning gains). Use of both measures accomplishes three things: First, it guards against lenient practice assessment, reducing the possibility that teachers who are not contributing substantially to student learning will achieve tenure or a higher career level. Second, it sends the message that both teaching practice and results are important. Third, it recognizes that measures have error, and to the extent that the two measures justify the same decision, it instills greater confidence in that decision.

As mentioned at the outset, legitimate concerns exist about the validity of using a value-added estimate of classroom productivity as the sole indicator of teacher performance for consequential decisions. A recent NRC study (Braun, Chudowsky, & Koenig, 2010) outlined many of these concerns. Temporal instability and large measurement errors, especially for smaller classes, are serious problems but can be addressed by combining estimates for multiple years, typically 2–3 (e.g., Koedel & Betts, 2009; Goldhaber & Hansen, 2008b). If multiple years of value-added estimates are available, it would seem logical to combine value-added measures with a teaching practice assessment for three human capital management decisions: (a) tenure, (b) career ladder or pay schedule movement, and (c) termination. For a tenure decision, a proficient level of practice, a specific level of value added, and perhaps a recommendation from the appropriate school leaders could be required. For movement to the highest level of a career ladder or knowledge- and skill-based pay schedule, it would seem reasonable to require the teacher to exhibit both exemplary practice and above-average value added. As for termination, it might make sense to target teachers who have consistently low value added for an intensive practice assessment that could lead to dismissal.

Three important questions need to be answered in combining practice assessment and value-added measures for human capital management decisions:[6]

1. Since teaching practice assessments and value-added indicators do not measure the same thing, how should they be combined? Should one of the two have primacy, should high levels of both be required, or should they be averaged?

2. How does one determine the value-added cutoffs for uses like tenure decision making? Unlike practice measures that are intended to be criterion-referenced, value-added measures are typically expressed in relation to an average classroom. There is no natural cutoff point that represents acceptable performance for tenure, outstanding performance for career progression, or poor performance warranting termination. This problem is most salient for tenure decisions. Although it may seem attractive to require a teacher to produce the average value added, there are drawbacks to this. Somewhere near half of the teachers will of necessity be below the average, and a state or district may not be able to dismiss nearly half of its new teachers. Supply may be insufficient, and additional resources unavailable for recruitment, selection, and induction. More conceptual and empirical work is needed to answer this question.

3. What should be done for teachers of nontested subjects? More test development is always an option, but in the short run it may be easier to adapt goal-setting approaches such as those used in Denver's compensation system (ProComp) and in Orange County, Florida, and Austin, Texas. The basic idea would be to have teachers and principals set classroom-specific goals for measurable student learning. Consistent attainment of goals could be considered equivalent to attaining above-average value added. Of course, safeguards would be needed to prevent gaming and leniency, but these may be less costly than the extensive program of test development needed to provide value-added estimates for all teachers.

The use of both value-added and judgment-based practice measures for human capital management decisions seems like a fruitful partnership. It could both produce better practice assessment systems and shed light on some of the puzzles of value-added estimates, such as their temporal instability. We recommend that districts that are serious about effective teaching assessment consider developing ways to make teaching assessments and value-added indicators work together.

[6] For additional discussion, see the appendix.

References

Braun, H., Chudowsky, N., & Koenig, J. (2010). Getting value out of value-added: Report of a workshop. Washington, DC: National Academies Press.

Bretz, R. D., Milkovich, G. T., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18*(2), 321–352.

Chester, M. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*(2), 32–41.

Chester, M. (2005). Making valid and consistent inferences about school effectiveness from multiple measures. *Educational Measurement: Issues and Practice*, *24*(4), 40–52.

Daley, G., & Kim, L. (2010). *A teacher evaluation system that works*. Santa Monica, CA: National Institute for Excellence in Teaching.

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.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.

Davis, D. R., Pool, J. E., & Mits-Cash, M. (2000). Issues in implementing a new teacher evaluation system in a large urban school district: Results of a qualitative field study. *Journal of Personnel Evaluation in Education, 14*(4), 285–306.

Dwyer, C. A. (1994). *Development of the knowledge base for the PRAXIS III: Classroom performance assessments assessment criteria*. Princeton, NJ: Educational Testing Service.

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.

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.

Goldhaber, D., & Hansen, M. (2008a). *Is it just a bad class? Assessing the stability of measured teacher performance* (CRPE Working Paper #2008-5). Retrieved from Center on Reinventing Public Education website: http://www.crpe.org/cs/crpe/download/csr_files/wp_crpe5_badclass_nov08.pdf

Goldhaber, D., & Hansen, M. (2008b). *Assessing the potential of using value-added estimates of teacher job performance for making tenure decisions* (Brief 3). Retrieved from National Center for Analysis of Longitudinal Data in Education Research website: http://www.urban.org/publications/1001265.html

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

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Erlbaum.

Hadden, D., & Pianta, R. (2006). MyTeachingPartner: An innovative model of professional development. *Young Children, 61*(2), 42–43.

Halverson, R., Kelley, C., & Kimball, S. (2004). Implementing teacher evaluation systems: How principals make sense of complex artifacts to shape local instructional practice. In W. Hoy & C. Miskel (Eds.), *Educational administration, policy, and reform: Research and measurement* (Research and theory in educational administration Vol. 3, pp. 153–188). Greenwich, CT: Information Age.

Heneman, H. G., III, & 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.

Heneman, H. G., III, & Milanowski, A. T. (2004) Alignment of human resource practices and teacher performance competency. *Peabody Journal of Education, 79*(4), 108–125

Heneman, H. G., III, & Milanowski, A. T. (2009, March). *Analyzing human resource practices alignment*. Retrieved from Strategic Management of Human Capital website: http://www.smhc-cpre.org/download/58/

Heneman, H. G., III, Milanowski, A., Kimball, S. M., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay (CPRE Policy Brief RB-45). Retrieved from Consortium for Policy Research in Education website: http://www.cpre.org/images/stories/cpre_pdfs/RB45.pdf

Harris, D. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard University Press.

Jaeger, R. M. (1998). Evaluating the psychometric qualities of the National Board for Professional Teaching Standards' assessments: A methodological accounting. *Journal of Personnel Evaluation in Education, 12*(2), 189–210.

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., & Milanowski, A. T. (2009). Examining teacher evaluation validity and leadership decision making within a standards-based evaluation system. *Educational Administration Quarterly, 45*(1), 34–70.

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.

Koedel, C., & Betts, J. R. (2009). *Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique* (WP 09-02). Retrieved from University of Missouri–Columbia Department of Economics Working Paper Series website: http://economics.missouri.edu/working-papers/koedelWP.shtml

Levy, P. E., & Williams, J. R. (2004). The social context of performance appraisal: A review and framework for the future. *Journal of Management, 30*(6), 881–905.

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.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). *The intertemporal stability of teacher effect estimates* (Working Paper 2008-22). Retrieved from National Center on Performance Incentives website: http://www.performanceincentives.org/data/files/news/PapersNews/McCaffrey_et_al_2008.pdf

Milanowski, A. T. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33–53.

Milanowski, A. T., Kimball, S. M., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. E. Schwartz, R. Rubenstein, & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (American Education Finance Association 2005 Yearbook, pp. 137–159). Larchmont, NY: Eye on Education.

Milanowski, A. T., & White, B. (2001). *Rating agreement in the 2000–2001 implementation of the Cincinnati Public Schools/Cincinnati Federation of Teachers teacher evaluation system*. Unpublished manuscript.

Moss, P. A. (2008). A critical review of the validity research agenda of the National Board for Professional Teaching Standards at the end of the first decade. In L. Ingvarson & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards* (pp. 257–312). Oxford, UK: Elsevier.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives*. Thousand Oaks, CA: Sage.

National Research Council. (2008). *Assessing accomplished teaching: Advanced-level certification programs* (M. D. Hakel, J. A. Koenig, & S. W. Elliott, Eds.). Washington, DC: National Academies Press.

Odden, A., & Wallace. M. (2008). *How to create world class teacher compensation*. St. Paul, MN: Freeload Press.

Pecheone, R. L., & Chung, R. R. (2007). *PACT technical report: Summary of validity and reliability studies for the 2003–04 pilot year*. Retrieved from the Performance Assessment for California Teachers website: http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf

Pianta, R., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*(2), 365–397.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System manual K–3*. Baltimore, MD: Paul H. Brookes.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2008). *The other 69%: Fairly rewarding the performance of teachers in nontested subjects and grades*. Retrieved from Center for Educator Compensation Reform website: http://cecr.ed.gov/guides/other69Percent.pdf

Reynolds, A. (1995). The knowledge base for beginning teachers: Education professionals' expectations versus research findings on learning to teach. *The Elementary School Journal, 95*(3),199–221.

Roberts, G. E. (1995). Municipal government performance appraisal system practices: Is the whole less than the sum of its parts? *Public Personnel Management, 24*(2),197–221.

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.

Ryan, J. (2002). Issues, strategies, and procedures for applying standards when multiple measures are employed. *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 289–315). Mahwah, NJ: Erlbaum.

Schacter, J., & Thum, Y. (2004). Paying for high- and low-quality teaching. *Economics of Education Review*, *23*(4), 411–430.

Schacter, J., & Thum, Y. (2005). TAPping into high quality teachers: Preliminary results from the Teacher Advancement Program Comprehensive School Reform. *School Effectiveness and School Improvement*, *16*(3), 327–353.

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

Whittaker, A., Synder, J., & Freeman, S. (2001). Restoring balance: A chronology of the development and uses of the California Standards for the Teaching Profession. *Teacher Education Quarterly*, *28*(1), 85–107.

Wilson, M., Hallam, P. J., Pecheone, R., & Moss, P. (2007). Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's *Beginning Educator Support and Training* program. Manuscript submitted for publication.

Yeh, S., & Ritter, J. (2009). The cost-effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance*, *34*(4), 426–451.

Table 1
Overview of Teaching Assessment Systems Reviewed

System	Original purpose	Extent of use	Theoretical perspective	Instrument structure	Specialization
CLASS	Research tool to assess the quality of early childhood (pre-K) teacher-student interactions. Additional versions developed for K–3 and middle/secondary use.	Pre-K version used for research purposes and in Head Start. K–3 and middle/secondary versions now used in several districts. Also used in U of VA's electronic teacher professional development program and ABCTE Distinguished Teacher certification.	Child developmental theory and research showing importance of quality of teacher-student interactions, including social and emotional functioning in the classroom.	Ten dimensions and 42 subdimensions grouped within 3 domains: emotional support, classroom organization, instructional support. Three performance levels explicitly defined by rubrics, but scoring augmentation within levels allows for 7 rating levels.	The 3 CLASS versions span pre-K to high school and apply to all experience levels and content areas. The version included in this study applies to K–3.
FAS	New teacher induction and mentoring.	Forty-six induction programs in CA; also used in NYC and Chicago induction programs.	Holistic view of teaching emphasizing content, student diversity and inclusion, differentiation of instruction, student engagement, and self-directed learning.	Six standards (domains): engaging students, environment for learning, understanding and organizing subject matter, planning instruction, assessing students, developing as a professional Six components/standard; 32 ratable dimensions. Rubrics define 5 performance levels.	Focus on novice teachers at all grade levels and content areas; higher rubric levels fit experienced teachers.
FFT–orig.	Formative tool for promoting conversations about good teaching. Suggested uses include self-assessment, induction and mentoring, peer coaching, and clinical supervision.	No data on extent of use available, but anecdotal evidence suggests use in some form in at least 200 districts of all sizes and types.	Intended to be a comprehensive representation of generic teaching activities applicable to almost all K–12 settings. Emphasizes aspects of constructivism.	Four domains: planning and preparation, classroom environment, instruction, professionalism. Twenty-two components (5–6 per domain) further divided into 66 elements; 66 ratable dimensions. Rubrics define 4 performance levels.	Intended to apply to all career and grade levels and all content areas.

Teaching Assessment for Teacher Human Capital Management

System	Original purpose	Extent of use	Theoretical perspective	Instrument structure	Specialization
Cincinnati TES	Teacher summative evaluation, including use in a career ladder program.	Single district; used primarily with newer teachers and teachers seeking teacher leadership positions.	Intended to drive instruction to fit district strategy by emphasizing student standards, engagement, and higher order thinking skills.	Fifteen rated standards (dimensions) grouped into 4 domains: planning and preparing, environment for learning, teaching for learning, professionalism. Rubrics define 4 levels of performance.	Intended to apply to all career and grade levels and all content areas.
NBPTS Math	Assessment of teaching practice as part of a voluntary certification system intended to recognize high-quality teachers.	NBPTS certification is supported, recognized, or rewarded in all 50 states and hundreds of districts. There are about 74,000 certified teachers.	Based on NBPTS's 5 core propositions: teachers (a) are committed to students and their learning; (b) know the subjects they teach and how to teach those subjects; (c) are responsible for managing and monitoring student learning; (d) think systematically about their practices and learn from experience; (e) are members of learning communities.	Four portfolio entries developed by teacher and 6 assessment center exercises requiring teacher constructed responses. Ten scores produced but do not directly represent proficiency levels on NBPTS standards. Exercises assess content knowledge; portfolio entries represent multiple standards. Rubrics define 4 levels, but intermediate scores can be given.	Experienced teachers (3+ years of experience) in middle school and early high school mathematics. Similar assessments used for 24 other certification areas.
PACT Math	New teacher initial licensure.	Thirty-two CA university and district teacher preparation programs.	"Authentic" assessment in place of paper-and-pencil tests; based on a plan-instruct-assess-reflect cycle, with special attention to subject-specific pedagogy and the teaching of ELLs.	Twelve guiding questions divided among 5 domains: planning, instruction, assessment, reflection, academic language. Rubrics define 4 performance levels.	New teachers to be licensed to teach middle and high school mathematics. Similar PACT assessments used for 25 other licensing areas, including elementary-grade generalist.

System	Original purpose	Extent of use	Theoretical perspective	Instrument structure	Specialization
PRAXIS III	Teacher licensure.	Used in OH and AR for licensure.	Development guided by assumption that (a) effective teaching requires both action and decision making, (b) learning is a process of active construction of knowledge, and (c) since good teaching depends on subject, no one teaching style is best for all contexts.	Nineteen dimensions grouped into 4 domains: organizing content knowledge, creating an environment for learning, teaching for learning, professionalism. Rubrics define 3 levels, but 2 intermediate levels can be scored, providing 5 possible ratings.	New teachers in all content areas and grade levels.
TAP	Formative and summative evaluation of teachers in TAP schools, including use in career ladder program and possible use for pay bonuses.	Impact on 10,000 teachers and 100,000 students in 2010–11.[a] Used in many TIF sites, some Q Comp sites in MN, and several districts in LA, NC, SC, and TX.	Eclectic mix of best practices drawn from various sources. Emphasizes high expectations, student engagement, teaching to standards, higher order thinking skills, use of assessment, and differentiation of instruction.	Twenty-six dimensions grouped into 4 domains: designing and planning instruction, learning environment, instruction, professional responsibilities. Three levels are defined, but raters can score in-between levels, providing 5 possible ratings.	Intended to apply to all career and grade levels and all content areas.

Note. CLASS = Classroom Assessment Scoring System, K–3 version. FAS = Formative Assessment System Continuum of Teacher Development. FFT–orig. = Framework for Teaching, original version. Cincinnati TES = Framework for Teaching, as adapted and implemented by Cincinnati Public Schools. NBPTS = National Board for Professional Teaching Standards mathematics/early adolescence assessment. PACT = Performance Assessment for California Teachers mathematics assessment. PRAXIS III = PRAXIS III teacher licensing performance assessment. TAP = Teacher Advancement Program. U of VA = University of Virginia. ABCTE = American Board for Certification of Teacher Excellence. ELLs = English language learners. TIF = Teacher Incentive Fund. Q Comp = Quality Compensation for Teachers.

[a] http://www.talentedteachers.org/action/action.taf?page=faq

Table 2

Coverage of Eight Important Teaching Competencies

System	Teaching competency							
	Content standards	**Use of formative assessment**	**Differentiation of instruction**	**Student engagement**	**Higher order thinking skills**	**Content knowledge & PCK**	**Personalized relationships with students**	**High expectations**
CLASS	0 0 0	0 0 0	2 20% 1	3 30% 1	3 30% 1	0 0 0	2 20% 1	1 10% 0
FAS	2 6% 0	2 6% 2	5 16% 2	5 16% 4	4 13% 2	2 6% 2	1 3% 0	2 6% 0
FFT–orig.	0 0 0	1 5% 0	4 18% 0	5 23% 1	3 14% 0	3 14% 1	1 5% 0	3 14% 1
Cincinnati TES	3 20% 1	1 7% 0	2 13% 0	3 20% 0	2 13% 1	3 20% 1	1 7% 1	1 7% 1
NBPTS Math	0 0 0	2 20% 0	1 10% 0	2 20% 0	2 20% 0	10 100% 6	0 0 0	2 20% 0
PACT Math	1 8% 0	2 17% 2	5 42% 1	1 8% 1	3 25% 1	6 46% 2	0 0 0	0 0 0
PRAXIS III	0 0 0	2 11% 0	2 11% 0	0 0 0	1 5% 0	1 5% 0	1 5% 1	1 5% 1
TAP	3 12% 1	1 4% 0	2 8% 0	3 12% 0	4 15% 2	1 4% 1	1 4% 0	2 8% 0

Note. CLASS = Classroom Assessment Scoring System, K–3 version. FAS = Formative Assessment System Continuum of Teacher Development. FFT–orig. = Framework for Teaching, original version. Cincinnati TES = Framework for Teaching, as adapted and implemented by Cincinnati Public Schools. NBPTS = National Board for Professional Teaching Standards mathematics/early adolescence assessment. PACT = Performance Assessment for California Teachers mathematics assessment. PRAXIS III = PRAXIS III teacher licensing performance assessment. TAP = Teacher Advancement Program. PCK = pedagogical content knowledge. The top number in each cell is the number of performance dimensions with rubrics that refer to the competency; the middle number is the percentage of these dimensions that refer to the competency; the bottom number is the number of scored performance dimensions that are predominantly based on the competency.

Table 3

Additional Common Teaching Competencies Found in Three or More Assessment Approaches

Teaching competency	PRAXIS III	FFT–orig.	Cincinnati TES	TAP	FAS	CLASS	PACT	NBPTS
Knowledge of students	■	■	■	■	□		○	○
Appropriate instructional goals	■	■	○	■	○		○	○
Communication of instructional goals	□		□	□	■	□		
Assessment aligned to goals	■	■	■	□			○	
Coherent instructional planning	□	■	□	■	■		□	○
Multiple assessment methods	□		□	□	■		□	
Assessment used to plan instruction	□	■	○	○	■		■	○
Positive relationships with students	■	■	■	□	■	■		
Fair, inclusive learning environment	■				■			○
Management of classroom procedures	○	■	■	○	■	■		
Use of instructional time	□	□	□	□	■	■		
Management of student behavior	■	■	■	■	■	■	○	
Student responsibility for behavior		□			□	○		
Physical organization of classroom	■	■		○	■	○		
Questioning/discussion techniques		■	■	■	□	□		○
Quality of feedback to students	○	■	■	■	○	■	■	○
Variety of instructional strategies	○		□		■	□		
Building on student experiences		□	□	○	■	□	□	○
Grouping of students		■	○	■	□			○
Student initiative in learning		□	○	○	■	□		○
Adaptation of lesson/plan		■	□	○	■	□	■	○
Reflection on practice	■	■	■	■	■		■	□
Communication with families	■	■	■		■			○
Cooperation with colleagues	■	■	■		■			○
Pursuit of professional development		■	■	■	□			○

Note. PRAXIS III = PRAXIS III teacher licensing performance assessment. FFT–orig. = Framework for Teaching, original version. Cincinnati TES = Framework for Teaching, as adapted and implemented by Cincinnati Public Schools. TAP = Teacher Advancement Program. FAS = Formative Assessment System Continuum of Teacher Development. CLASS = Classroom Assessment Scoring System, K–3 version. PACT = Performance Assessment for California Teachers mathematics assessment. NBPTS = National Board for Professional Teaching Standards mathematics/early adolescence assessment. ■ = Contains a performance dimension primarily based on this competency. □ = This competency is a major consideration in scoring on one or more dimensions. ○ = This competency is mentioned in the rubrics.

Table 4
Summary of Assessment Processes

System	Data collection methods	No. observations/occasions	Rubrics	Customization
CLASS	Classroom observation.	Minimum of four 20-minute observation cycles in a 2-hour period required, with six preferred. No requirement for multiple days.	Five-level rating scale with behavioral anchors at Levels 1, 3, & 5 for 10 dimensions. Four to five subdimensions within each dimension (42 total; not rated separately).	Versions for pre-K & Grades 4–12 available.
FAS	Classroom observation, artifact collection.	Three visits suggested, spread out over several days or a week, twice a year.	Six-level rating scale with behavioral anchors at all levels for 32 dimensions within 6 domains.	None.
FFT–orig.	Classroom observation, artifact collection, teacher written reflections.	Not specified.	Four-level rating scales with behavioral anchors at each level for 66 elements, grouped within 22 components grouped within 4 domains.	Users encouraged to customize rubrics. Has been customized for support professionals.
Cincinnati TES	Classroom observation, artifact collection, teacher written reflections.	Four to six full-period observations, depending on purpose & teacher experience.	Four-level rating scales with behavioral anchors at each level, for 15 dimensions. Most dimensions have subdimensions (32 in all; not separately rated).	Standard wording customized & interpretation notes added for some content areas.
NBPTS	Videotape & performance tasks including commentary on videos, reflections; demonstrations of content knowledge.	Two 15-minute videos; 6 performance tasks related to content knowledge (resemble constructed-response tests).	Four performance levels defined based on quality of evidence for each of 10 performance tasks. Holistic rubric for each task includes narrative with examples of 4 performance levels. Intermediate levels can be scored.	Assessments customized by age & content in 25 certification areas.
PACT	Videotape & performance tasks (e.g., commenting on videos, designing lesson plan, writing reflections).	One or two video clips at discretion of teacher.	Four-level rating scale with behavioral anchors at each level for 12 dimensions.	Assessments customized to content in 5 elementary & 18 secondary certification areas via wording changes in standard rubrics.
PRAXIS III	Classroom observation, teacher interview.	Two/year recommended, each by a different assessor.	Five-level rating scale with behavioral anchors at low, intermediate, & high levels for 19 dimensions.	None.
TAP	Classroom observation, artifact collection.	Four to six/year; assessors encouraged to do more if key evidence not observed.	Four-level rating scales with behavioral anchors at each level, for 19 dimensions grouped within 3 domains. The additional 7 dimensions in 4th domain (4 of which apply only to teacher leaders), have 3-level labeled rating scales rather than rubrics.	None for subject or level. Local users encouraged to customize scales for professionalism domain.

Note. CLASS = Classroom Assessment Scoring System, K–3 version. FAS = Formative Assessment System Continuum of Teacher Development. FFT–orig. = Framework for Teaching, original version. Cincinnati TES = Framework for Teaching, as adapted and implemented by Cincinnati Public Schools. NBPTS = National Board for Professional Teaching Standards mathematics/early adolescence assessment. PACT = Performance Assessment for California Teachers mathematics assessment. PRAXIS III = PRAXIS III teacher licensing performance assessment. TAP = Teacher Advancement Program.

Table 5
Suggested Data Collection and Assessment Methods for Human Capital Management Uses

HCM use	Suggested data collection and assessment methods
Initial selection	• Interview question bank with multiple questions and suggested rating scales based on the competencies • Demonstration lesson rated using observational protocol • Reference check protocol based on key competencies
Induction and mentoring	• Observation tool focusing on competencies being developed by induction and mentoring program, including pre- and post-observation conferences
Professional development	• Videos of lessons and artifacts from an instructional unit, to be assessed off-site by an expert[b]
Performance management	• Teacher evaluation tool based on multiple live observations by school leaders, some prearranged so assessor can see how well teacher is implementing suggestions • Walk-through tools focusing on judgments by school leaders of whether 2–4 readily observable competencies are being displayed
Compensation[a]	• Performance assessment based on instructional units, scored by external assessors[c]

Note. HCM = human capital management.
[a]For example, a career ladder or knowledge- and skill-based compensation system.
[b]This expert could also provide feedback and coaching. The work samples could be used by teachers to prepare for consequential assessments or by lesson study groups.
[c]This assessment could require (a) at least a proficient performance evaluation rating by school administrators to be eligible for movement and (b) evidence of student learning for progression to highest levels. Additional considerations on using teaching assessments as part of a teacher compensation system can be found in Heneman, Milanowski, Kimball, and Odden (2006).

Appendix

Using Value-Added and Instructional Practice Measures Together for Human Capital Management Decisions

As discussed in the main body of this paper, using value-added estimates of classroom productivity together with assessments of teaching practice as the basis for human capital management decisions is an attractive possibility. Assessment experts and policy researchers (e.g., Guion, 1998, Chapter 14; Harris, 2011) have recommended the use of multiple measures when high-stakes decisions are to be made, and many human capital management decisions can be high-stakes for teachers. However, attractive as the idea of multiple measures seems, several issues need to be thought through to gain the potential benefits.

The first is whether the value-added measure and the teaching assessment measure should be combined into one overall measure of performance. This approach seems logical given the way we usually think about the advantages of multiple measures. They are typically advocated on the basis that both measures contain error, and combining them "averages out" the error to produce a more reliable measure. However, there are both conceptual and technical reasons why this rationale does not apply well in the case of value-added and teaching practice measures.

Fundamentally, the problem is that value-added and teaching assessment scores do not measure the same thing. Classroom value-added estimates represent the average difference in students' achievement from an expected level or from the average growth in achievement, depending on the model used. Teaching practice assessment scores measure how well teachers can or do exhibit desired instructional behaviors or skills. Classroom value added is intended to be a measure of average student learning, whereas teaching assessments are measures of teacher practice, which is only one of several causes of student learning. Thus, at the construct level the two indicators are not equivalent. They should be correlated, but this correlation is not likely to be as high as one would expect between reliable measures of the same thing (e.g., .7–.9). Further, the predominant sources of measurement error in teaching assessment scores (assessor effects, occasion-of-observation effects) are not the same as those for value-added measures (student sampling error, misalignment of test content with enacted curriculum) and thus will not average or cancel out when the two are combined. In this case, the reliability of a composite is close to the average reliability of the two components (Ryan, 2002; Chester, 2003) rather than a substantial improvement over the reliability of either.

A related technical consideration is that the measures have different metrics, so one cannot simply add or average them. One might consider standardizing them before adding or averaging, but this would change the interpretation of the instructional practice score. Such scores are intended to be, and are usually treated as, measures of how well practice reflects a criterion standard of performance independent of any particular teacher. This standard of performance is intended to be defined by the rubric or rating scale. Standardizing these scores makes them into a measure of performance relative to the average, thus obscuring a teacher's position with respect to practice standards (e.g., at, above, or below proficiency). This lowers the formative utility of the scores.

Moreover, adding or averaging the two scores implies acceptance of a compensatory model, even if the component scores are weighted differently, and it is unclear whether

stakeholders would be comfortable viewing value-added and teaching practice measures in this way. Some might be willing to say that a teacher could make up for a low level of teaching practice with a high level of value added, but the reverse would likely be considered problematic.

The above arguments support treating the teaching assessment score and value-added estimate as separate performance measures. How, then, can they be used together for various human capital management decisions?

One possibility is what measurement experts call a *conjoint model*, in which minimum scores on both measures are needed to trigger some decision (Ryan, 2002; Chester, 2003). This is a natural model to use for a tenure decision. For example, a district might require a teacher to demonstrate (a) a proficient level on the practice assessment by the third or fourth year and (b) an average value added above some level based on 3 years of estimates. This approach would be aimed at ensuring that the teacher can both practice according to the standards and effectively facilitate student learning.

The second issue to be considered in using both value-added and teaching practice measures is how to set the minimum value-added cutoff point for consequential decisions such as tenure. Since value-added estimates are relative to the other teachers in the state or district teaching force, there is no natural cutoff point that represents acceptable performance for tenure. Requiring "average" value added sounds attractive because a teacher with average value added could be interpreted as producing the expected amount of growth in student achievement. However, this approach could result in more teachers failing to achieve tenure than can be effectively replaced. If the value-added comparison was made to an average based on the whole teaching force (actually, all others teaching the tested subjects), one would likely find *more* than half of the teachers being considered for tenure below average, simply because less experienced teachers generally have lower value added than those with 5 or more years of experience. If the value-added comparison was made within the group of new teachers coming up for tenure, using the average as a cutoff would eliminate about half of these teachers. This approach might lead to a more manageable number of teachers to replace while also being a fairer "apples to apples" comparison.

Another possibility would be to set the minimum based on a confidence interval below the average value added of the teachers coming up for tenure in any year. For example, the minimum could be the lower limit of the conventional 95% confidence interval. Narrower confidence intervals could be used if the decision makers were more concerned about false positives (granting tenure to teachers who later prove to be less effective than predicted) than false negatives (losing teachers who turn out to be more effective). For teachers responsible for multiple tested subjects, this minimum would be required for each subject. Yet another approach would be to calibrate value added in terms of the gains needed to move students to state proficiency standards. One could use value-added estimates to develop expected trajectories for students and then require value-added levels sufficient to maintain these trajectories. In any case, districts would want to estimate how many teachers would be terminated under this approach and consider whether the additional terminations could be replaced without lowering the hiring bar too far, given the supply of new teachers.

Gordon, Kane, and Staiger (2006) showed that new teachers being considered for tenure who were in the bottom quartile of the value-added distribution would improve the overall quality of a district's teaching staff. Using data from Los Angeles, Gordon and his colleagues determined that such a policy would yield a net increase in average value added despite the fact that the increase would be partially offset by the lower value added of the greater number of novice teachers hired as replacements. A district following this policy would likely have a higher proportion of novice teachers who would have on average lower value added than that of the teachers they replace because first-year teachers tend to have lower value added than second-year teachers. This paper illustrates the cost-benefit analysis districts may want to undertake when deciding on a value-added cutoff. Districts will also need to consider whether the additional costs of recruiting and inducting more novice teachers can be met and whether the pool of potential new teachers is big enough to meet the increased demand.[7]

For a termination decision about a tenured teacher, it would likely be more acceptable to require both low value added and low teaching performance assessment scores. This approach would require setting both a value-added minimum and a practice score minimum. The latter is provided by the design of the assessment rubric; districts are likely to simply require teachers to be rated proficient on all practice dimensions. It is harder to set a value-added minimum, because value-added estimates reflect performance relative to the group of classrooms being measured. Unless the group is quite large, any teacher's classroom value added will also be sensitive to how well other classrooms do. This fact—coupled with the arbitrariness inherent in setting a threshold (e.g., the bottom decile or quintile of the distribution)—would make it easy to argue that poor value added is not solely the teacher's responsibility.

A more promising approach might be to use consistently low value added as an initial signal that a teacher needs to be reviewed, no matter how high the practice assessment scores. Teachers with very low value added—say, in the bottom 10% of a 3-year average or in the bottom 20% 3 years in a row—would have their practice reviewed by an evaluator from outside the school.[8] If this evaluator found that practice was below the proficient level, the teacher would be given a year to improve. 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. The policy might also require the teacher to have classroom value-added results above the bottom 20% for the next 2 years in order to avoid another outside evaluation.

The above scenario suggests a 3-year cycle for teacher evaluation combining value-added and practice assessment. That is, tenured teachers' practice and value added would be reviewed every 3 years, and consequential decisions made based on the results of that 3-year review.

[7] Yeh and Ritter (2009) argued that high replacement costs and costs associated with expanding the pool of new teachers would be significant. However, comparing the Gordon et al. (2006) proposal to nine other interventions, Yeh and Ritter concluded that replacing new teachers with low value added was a more cost-effective approach than all but three of the alternatives considered..

[8] If there was a substantial overall correlation between value added and practice assessment ratings in a district, we would expect to find the teacher with consistently low value added also delivering below-average instruction and so to suspect their prior ratings might have been due to leniency. This is why an evaluator from outside the building would be desirable.

For a career ladder or knowledge- and skill-based pay system, using both teaching assessment scores and value added again seems promising. To move to the top level of such a system, a teacher could be required to have the highest scores on the teaching practice assessment and a high level of value added. Again, multiple years of value-added estimates would need to be used, and the cutoff determined (e.g., the top quartile or quintile or a point outside the upper confidence level around the mean value added). In setting this cutoff, a district would want to review the distribution and reliability of classroom value-added estimates, as well as the funding available to pay for salary increases. Table A1[9] shows an example of how practice assessment and value-added measures could be combined in a knowledge- and skill-based pay structure.

The value-added requirement for the Career level would be the complement of the requirement that a tenured teacher stay out of the bottom 20% of the value-added distribution to avoid outside review, as discussed above. Note that there are two ways teachers could qualify for the Accomplished level: (a) if their instructional practice was above proficient on the majority of practice dimensions and their value added was never in the bottom 20%; or (b) if their value added was consistently in the top 50% and their instructional practice was rated proficient on all dimensions. This approach would allow teachers with good practice to be recognized, even in subject areas for which value added is not as stable. It would also allow those who might not want to follow the orthodoxy of district practice standards to be recognized, as long as they are consistently producing better-than-average student achievement.

It might appear at first that Option 2 at the Accomplished level sets a low bar (being in the top 50% of the value-added distribution of all teachers at the Career level and above in 2 consecutive years). However, because of the considerable year-to-year variation in value added, far less than 50% of teachers would meet this criterion. Using data from Goldhaber and Hansen (2008b) and McCaffrey et al. (2008) on the percentages of teachers who change or remain in value-added quintiles over 2 years, we estimate that somewhere between 20% and 30% of teachers would meet this requirement. Given 2 years of value-added data, a district could be reasonably sure that these teachers were making an above-average contribution to student achievement.

Performance pay can also be implemented without directly combining value-added and teaching assessment results. For example, if performance pay is provided by one-time bonuses, the stakes are lower and so is the concern about false positives and negatives. In this case, a natural model is the performance scorecard that simply reports the results of separate measures of performance (like a student's report card) and associates a bonus amount with achievement of performance goals on each measure. This is a simple approach that is also easy to understand and allows teachers to be recognized for practice, value-added results, or both.[10] This is the general approach taken to determine performance bonuses in the TAP model. Table A2 shows a generic example of a scorecard.

[9] Tables A1 and A2 follow the appendix.

[10] Of course, this approach is more justified if there is a positive correlation between practice assessment scores and value added, because then the incentive provided to improve practice scores would also promote an average improvement in value-added classroom productivity.

In the example scorecard, practice assessment and student achievement are weighted equally, and each has an associated dollar amount that is earned by meeting a performance target (the score needed to earn the bonus). The teacher's bonus is simply the sum of the bonus amounts earned by meeting each separate performance target. Note that this is a compensatory model because a teacher can earn a bonus based on value added if teaching practice is below the target or vice versa.

In deciding on the weights to be placed on practice assessment and value added, incentive designers would probably want to consider both the reliability of the measures and the acceptability to stakeholders. For an annual bonus, value-added estimates from one year would be acceptable, but it would still be useful in establishing the performance target and weight to consider the possibility of misclassification due to student sampling. Similarly, one would want to consider the interobserver agreement of practice scores, the distribution of those scores, and their overall correlation with value-added estimates. The threshold for receiving a bonus based on practice assessment would need to be set high enough so that the state or district could be reasonably certain that teachers scoring at that level would be exhibiting superior performance and that their classrooms would have a high probability of showing above-average value added.

Table A1

Combining Value-Added and Practice Measures to Define Career Levels in a Knowledge- and Skill-Based Pay Structure

Career level	Practice assessment[a]	Value added[b]	Typical length of time at this level
Entry	Initial license	n/a	First-year level: Teachers have 2 years to reach next level or be terminated.
Developing	Rated at least *basic* (Level 2) on all practice dimensions	n/a	Teachers spend 2–4 years at this level; those who do not move to next level at end of 4th year are terminated.
Career	Rated *proficient* (Level 3) on all practice dimensions	In top 80% of all teachers beyond *Developing* level for 3 consecutive prior years	Tenure level: Teachers remain at this level indefinitely as long as they maintain required teaching practice and value-added performance.
Accomplished	*Option 1:* Rated *distinguished* (Level 4) on majority of practice dimensions and *proficient* on others	In top 80% of all teachers beyond *Developing* level for 3 consecutive prior years	Teachers remain at this level indefinitely as long as they maintain required teaching practice and value-added performance. Teachers reviewed every 3 years.
	Option 2: Rated *proficient* (Level 3) on all practice dimensions	In top 50% of all teachers beyond *Developing* level for 2 consecutive prior years	
Advanced	Rated *distinguished* (Level 4) on 75% of all practice dimensions	In top 50% of all teachers beyond *Developing* level for 2 consecutive prior years	Same as *Accomplished* level.

Note. Adapted from concepts presented in *How to Create World Class Teacher Compensation* by A. Odden and M. Wallace, 2008 (Freeload Press).

[a]Assumes an assessment system with multiple performance dimensions and four defined performance levels, such as the FFT.

[b]The value-added requirement is *in addition to* the practice assessment requirement at each level.

Table A2
Performance Scorecard Used to Determine Bonus Payments for Hypothetical Teacher

Performance dimension	Measure	Weight	Performance target for bonus	Teacher *A*'s score	Bonus amount
Teaching practice	5-dimension, 4-level practice assessment	50%	Average score of 3.5 (out of 4) across 5 performance dimensions	3.6	$1,250
Student achievement	Classroom value added	25%	Top 25% of classrooms in reading	80th pctile	$625
		25%	Top 25% of classrooms in math	55th pctile	$0
				Total attained	$1,875
				Total possible	$2,500