SMHC
STRATEGIC MANAGEMENT OF HUMAN CAPITAL

ALLAN R. ODDEN AND JAMES A. KELLY
CO-DIRECTORS

REVIEW OF TEACHING PERFORMANCE ASSESSMENTS FOR USE IN HUMAN CAPITAL MANAGEMENT

Anthony T. Milanowski
Corresponding Author
Consortium for Policy Research in Education
Wisconsin Center for Education Research
University of Wisconsin-Madison
1025 W. Johnson St.
Madison, WI 53706
(608) 262-9872
amilanow@.wisc.edu

Herbert G. Heneman III
Graduate School of Business and
Consortium for Policy Research in Education
University of Wisconsin-Madison

Steven M. Kimball
Consortium for Policy Research in Education
Wisconsin Center for Education Research
University of Wisconsin-Madison

WORKING PAPER
August 2009

The work described in this paper was supported by the Ford Foundation. The opinions expressed are those of the authors and do not necessarily reflect the view of the Ford Foundation, the institutional partners of CPRE, or the Wisconsin Center for Education Research. This is a preliminary summary of work in progress. Comments and suggestions welcome.

Why Measure Teaching Performance?

Teacher performance in the classroom is the lifeblood of the educational enterprise. Teachers weave a combination of their knowledge, skills and abilities into specific performance competencies that become drivers of student learning and achievement. Thus systems or processes for measuring teaching have been increasingly recognized as an important part of the instructional improvement puzzle. The assessment of teaching performance is a critical part of any attempt to develop a coherent system for the strategic management of teacher human capital. Such assessments provide both a measure of how well we are achieving such management, and are part of the management system itself.

The teacher performance competencies that form the basis for measuring teaching performance can be identified through a combination of research and expert judgment. Moreover, the competencies can be cast into a formal teacher performance competency model. The performance competency model depicts a broad representation of the teacher behaviors deemed desirable for effective classroom instruction. Usually the competency model specifies key performance domains, and within each domain are the specific behaviors. For example, virtually any teacher competency model will have a domain pertaining to actual delivery of classroom instruction, and within that domain will be numerous instructional behaviors such as "using assessments in instruction". Examples of other common domains are instructional planning, classroom management, interactions with others (staff, parents), and professionalism.

Desired teacher behaviors, whether or not they are cast into a formal competency model, 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 (e.g., ratings) and qualitative (e.g., written comments to support a rating score or suggest areas for improvement) information. In turn, the information 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 (pre-service and after-hire)
4. Mentoring
5. Professional development
6. Performance management
7. Compensation

These HR practice areas can be aligned to the teacher performance competencies in order to help the district acquire, develop, and retain a competent teacher workforce. Such alignment requires that the competencies be embedded within the HR practices, such as having professional development activities that focus on improvement of the desired competencies. This type of alignment is referred to as vertical alignment. Another type of alignment is referred to horizontal alignment, in which various HR practices are supportive of each other in their competency

emphasis. A teacher evaluation system that provides information about teacher competency deficiencies and is used as input to teachers' specific professional development plans is an example of horizontal alignment between teacher evaluation and professional development.

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 the 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 to) the performance competencies and assessment. Because the HR practices themselves all have a common foundation, they are meshed with each other (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
Human Resource Alignment and Performance Assessment

Teaching performance assessment can be used in each of the seven HR practices. For example, some form of teaching assessment could be part of a district's selection process. Formative teaching assessment could be part of 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 all these programs are based on the same competencies, which are then assessed in a consistent manner, they 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 way of thinking about and talking about good teaching. This common language in turn is central to developing a culture of high instructional performance.

There are those who might wonder why it is necessary to measure instructional practice when value-added technology lets us measure teacher and school contributions to student achievement. While value-added is an important tool, it is not sufficient on its own as an indicator of teaching performance for use in strategic human capital management. There are several reasons why this is so. First, value-added can only be used to measure the performance of teachers of regularly-tested subjects. The Center of Educator Compensation Reform, created by the US Department of Education to support the implementation of teacher performance incentives based on outcome measures, estimates that the student test data needed for value-added analysis is not available for about 69% of the nation's teachers (Prince et al., 2008). Second, value-added measures can usually only reliably distinguish between the highest and lowest performers, say the top versus the bottom 20%. Even for these teachers, value-added estimates often differ substantially from year to year (Goldhaber and Hanson, 2008; McCaffery et al., 2008). Third, value-added estimates show only how well a teacher's students are doing, on average, compared to other teachers' similar students. But if improving value-added requires improving instruction, teachers and administrators need to know what specific instructional practices need to be changed. Lastly, while value-added can be used to evaluate, retain, and pay teachers, it is much less clear how value-added estimates could be used to recruit and select new teachers, to identify specific professional development needs, or to structure induction and mentoring programs. Thus we believe that value-added and behavioral measures of teaching performance will need to work together for the foreseeable future. We discuss some ways value-added and teaching practice assessments can be used together in the final section of this paper.

Purpose of Study

The major goal of this study is to review the current state of the art in teaching assessment by examining a sample of assessment systems, then to develop a "specification" for a state-of the art performance assessment system to be used for HCM functions. This specification could be a stimulus and guidepost for working on a coherent instructional vision and methods to assess how well actual instruction reflects the vision. The results could also help states or 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 a first look at a "specification" for a high quality, multi-use assessment system, and a preliminary roadmap for developing such a system.

The Method of This Study

In order to inform our thinking toward the development of a framework for a comprehensive teaching assessment system that could be used for teacher human capital management, we began

with a review of several prominent assessment systems. We chose seven that seemed to incorporate best practices in assessment technique, cover a range of teacher performance competencies, and represent a variety of different approaches and uses. We limited our review to systems that are currently in use by states or 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 what the key competencies for teaching are and how to reliably measure 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 HCM purposes. We chose seven systems for this study.

- The PRAXIS III teacher licensing 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 developed by the New Teacher Center at the University of California at Santa Cruz

- The Framework for Teaching, originally developed by Charlotte Danielson, including the variant adapted and implemented by the Cincinnati Public Schools

- 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]

- The Classroom Assessment Scoring System (CLASS) developed by Robert Pianta and his colleagues at the Center for Advanced Study in Teaching and Learning at the University of Virginia.

CLASS, PACT, and the National Board assessments have different versions. We chose to examine the K-3 version of CLASS, the PACT mathematics assessment, and the Early Adolescent mathematics assessment of the National Board. We chose mathematics for both the PACT and National Board assessments in order to compare and contrast the 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.

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

Table 1 provides a brief overview of each of the systems we reviewed, including information on its original purpose and current uses, the theoretical perspective on which each is based, and the basic structure of the instrumentation.

<Insert Table 1 Here>

To begin the study, we obtained documents describing the systems, including descriptions of the performance dimensions and rubrics, training offered, and uses proposed. It is important to note that two systems, PACT and the National Board assessment, actually use somewhat different standards and rating scales for different teaching content areas and/or student age groups. For these two systems, 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 dimensions related to use in HCM activities. These were:

- Similarities and differences in underlying competency models
- Similarities and differences in assessment procedures
- Research related to reliability and validity

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

This paper is an initial draft of the findings from the study. As the study continues, some of the details of the conclusions may change. Additional comparison areas will also be added, including comparisons of administrative feasibility and likely teacher/administrator acceptance.

Study Results to Date

Two Basic Assessment Approaches

One major division among the systems is based on the primary method of collecting evidence of performance competency. The NBPTS assessment and PACT are both performance assessments in the technical sense. That is, both ask teachers to complete defined performance tasks that are designed to allow teachers to demonstrate specific knowledge, skills, and abilities. It is notable that both systems also have separate standards and rubrics for different subject areas and/or grade levels. These similarities are not accidental, stemming from both purpose and design. Both of these assessments were designed for certification, and the assessments therefore concentrate on whether teachers have the skills to perform, rather than their typical performance is therefore a one-time event. PACT developers also told us that they used the National Board assessment process as a general model in developing both PACT and its ancestor, the BEST licensing assessment formerly used in Connecticut.

The other systems are primarily based on live observation in classrooms. Within this group, PRAXIS III, Framework for Teaching (the original and Cincinnati version), and TAP have a degree of family resemblance. The development of PRAXIS III influenced the development of the Framework, which in turn influenced the development of the TAP model. Charlotte Danielson, the author of the Framework, has been involved with both ETS in PRAXIS III development and with NIET. 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. There are clearly concepts that are found in all three systems, and sometimes even similar phrases are used in the rubrics. It could almost be said that the TAP system is the latest step in a development that began with PRAXIS III.

While the FAS Continuum is not genetically related to this family, it does have some similarities with the Framework and PRAXIS III, both with respect to content and in being based on 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 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.

Similarities and Differences in Underlying Competency Model Content

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 be 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 seven 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 way involved counting how many performance dimensions there were in each system that referenced eight important competencies that underlie the kind of teaching that is often regarded as most likely to improve student achievement:

1. Attention to Student Standards
2. Use of Formative Assessment to Guide Instruction
3. Differentiation of Instruction
4. Engaging Students
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
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 Framework for Teaching has four domains, 22 components, and 66 elements. In this case, we decided to consider the 22 components as the performance dimensions, because it is unlikely that a district would want to score teaching on all the 66 elements. Scoring 22 components seems adequate for most uses. Similarly, though the Cincinnati system has rubrics that define performance levels for 32 separate dimensions of teaching, since these are grouped into 15 standards and teachers are rated only on these 15, we chose the 15 standards as dimensions. We chose the ten dimensions of CLASS that are normally scored, even though the rubrics might be used to scoring of 32 aspects of teaching. Like the Framework, the FAS Continuum does not specify the level which is scored. While the FAS Continuum has 6 standards, this seemed like too few, so we treated each of the 32 rubric strands as a performance dimension. The other systems do not have this sort of structure. TAP has 26 rated dimensions, PRAXIS III 19, and PACT 12.

Including the NBPTS assessments in these comparisons was complicated because the National Board standards differ by area of certification. Though themes carry through different areas, the standards differ in name and content across certification areas. (In contrast, PACT uses the same dimensions but words some of them differently for different subjects.) Another complication is that the National Board standards are not directly rated or assessed. There are no rubrics that define 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 a score on each exercise, not each standard, we decided to treat the four portfolio entries and six assessment center exercises as the performance dimensions and use the rubrics for scoring these 10 dimensions in our comparison of the competency models.

Before we discuss the results of this analysis, we need to make clear that these are preliminary results. We are in the process of assessing the reliability of the decisions we made about which systems cover which aspects of teaching.

Table 2 attempts to portray each system's coverage of the eight teaching competencies listed above. In each cell, there are three numbers that attempt to portray how the system covers each of the eight. The first number in each cell is the number of performance dimensions that refer to the competency in any substantial way in each set of scoring rubrics. This indicates the broad coverage of the competency. The second is the percentage of the total number of dimensions that refer to each competency. This indicates the relative emphasis the system places on each. The third is the number of dimensions that are *predominantly* focused on that competency. This shows whether a teacher is likely to receive a score or rating that primarily reflects the competency. Note that both the original version of the Framework for Teaching (Danielson, 1996), and Cincinnati Public Schools' adaptation of the Framework are included in the table.

<Insert Table 2 Here>

There are several conclusions that can be drawn from Table 2. First, it is clear that the National Board assessment puts the most emphasis on content knowledge and pedagogical

content knowledge. Both the portfolio entries and the assessment center exercises assess these constructs. The assessment center exercises in particular require substantial mathematics content knowledge, and are essentially a test of knowledge of specific areas of mathematics. PACT also emphasized these constructs in three of the 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 five of its 12 dimensions. All of the other systems reference differentiation in at least one dimension. CLASS puts the most emphasis on personalized relationships with students, likely due to its K-3 focus and theoretical base. 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 Framework, the National Board assessment, or PRAXIS III. In contrast, the Cincinnati adaptation of the Framework 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 tried to assess the similarity of the content of the competency models was to lay out the rubrics of the systems side by side and identify those 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 aspect, which contained dimensions for which the aspect was a major but not dominant part of the associated rubric, and which systems mentioned that aspect as one of several factors in a rubric. Table 3 lists these aspects, less those that repeated the eight constructs from Table 2. When considering this table, it is important to note that when an aspect or concept was not explicitly mentioned in a set of rubrics or dimensions, we did not count it, though it could be argued to be implicit in the rubric.[2]

<Insert Table 3 Here>

Table 3 also shows that there is a substantial amount of common content across the different approaches. Almost all of the systems we reviewed have substantial content related to the following teaching competencies:

[2] In the case of CLASS, some dimensions include sub-dimensions which 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 only one or two sub-dimensions address it.

- Knowledge of students
- Setting instructional goals
- Planning instruction
- Plan/lesson adjustment
- Building on student interests and experiences
- Managing classroom procedures and use of instructional time
- Managing student behavior
- Providing feedback to students
- Reflection 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 NB 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 Framework for Teaching, the Cincinnati system, TAP, and the FAS Continuum cover generic pedagogy in more detail. These systems, especially TAP and FAS, contain rich 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. A notable 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 like a promising addition to teaching assessment, given the importance of language for all academic subjects.

Comparison of the TAP and the Cincinnati systems with the Framework for Teaching illustrates a potential trade-off between comprehensiveness and focus. The Framework is perhaps the most comprehensive in what is assessed, while TAP and the Cincinnati version put more relative emphasis on specific aspects of the instructional vision thought to be important in the strategy for improving student achievement, such as alignment of instruction to student academic standards in Cincinnati and higher order thinking skills in TAP. Both also zero in more explicitly than the Framework on conceptual understanding and differentiation. Perhaps as the price for comprehensiveness, the Framework can be schematic is some areas. (Note that Danielson suggests users add to the Framework to meet local needs.) The FAS Continuum is an interesting balance between comprehensiveness and specificity, with rubrics that are often more detailed then those in the original Framework. However, because of the developers' desire to present teaching as holistic, some dimension descriptors and rubric language can seem somewhat redundant. Both the FAS Continuum and the Framework put greater emphasis on student autonomy and responsibility. The Framework uses these ideas to define higher levels of practice in nine of the performance dimensions (components), while the FAS Continuum covers this concept in five of the performance dimensions. CLASS also covers it in two rubric strands of one dimension.

Summarizing the content comparisons made 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 Framework for Teaching and the FAS Continuum are the most comprehensive in terms of teaching behavior. TAP, and the Cincinnati system are less comprehensive, but arguably more focused on specific aspects of instruction likely to influence student achievement. They may represent a worthwhile trade-off of breadth for depth and focus 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 National Board 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 is notable for its customization around subject mater while retaining similar performance dimensions across assessment areas.

Similarities and Differences in Assessment Processes

In considering the similarities and differences among the assessment processes, 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 stem form the purposes of the systems, but others are related to the assessment technology employed. Because NBPTS and PACT present teachers with a standardized set of tasks or prompts, they can more easily be structured to get at specific and often complex behaviors directly. In contrast, systems based on observation typically take as evidence whatever is observed (or whatever is shown by the artifacts collected) and so may not collect as much information on some competencies or areas of performance. But because the PACT and NBPTS assessments use standardized tasks, they also provide teachers with the opportunity to prepare the response. This means that what assessors may see is not typical, but peak performance (that is, more or less about the best the teacher can do[3]). Given the purposes the NBPTS and PACT assessments are used for, this is appropriate. Observation-based systems, even when supplemented by artifacts like lesson plans and student work, are more likely to capture typical performance, albeit at the expense of potentially missing evidence relevant to some performance dimensions unless many observations are used. This is because teacher behavior tends to vary substantially over a day, week, or year. This needs to be taken into account when assessments based on passive observation are used for consequential decisions.

The greater emphasis of the NBPTS and PACT assessments on content and pedagogical content knowledge noted above is also reflected their 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 systems use the same basic method of observation and rubrics across all content areas.

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

Despite the major divide between the NBPTS and PACT assessments and the others, there are several commonalities among the systems. Perhaps the most obvious is that all have a set of specific, separate performance dimensions or standards. While most of the systems' documentation explicitly recognizes that teaching is a complex activity in which several competencies are used in an intertwined way to produce proficient or accomplished teaching, all do break down teaching competence 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 judgments required are broken down and focused on specific types of evidence. This promotes reliable scoring. A system requiring just a global rating of performance or skill would allow each assessor to weigh the various aspects of teaching differently, and apply very different models of scoring. For example, one assessor may believe that no teacher can be competent without nearly perfect classroom management, while another may believe that a high level of student engagement can make up for imperfect classroom management.

While all 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 equal, the greater the number of dimensions, the more breadth is possible and the more fine-grained the analysis of teaching. A finer-grained system provides more information on desired performance to teachers, and allows the system to reflect individuals' strengths and weaknesses. But this comes at the expense of more complex scoring. The Framework for Teaching has the most potential dimensions, with a total of 66 elements that could be rated. One reason is that the Framework is the most comprehensive competency model, including several types of competency not found in the others (e.g., supervision of paraprofessionals and volunteers). It also analyzes the generic process of teaching down to a relatively fine level, at the expense of including many potential dimensions and perhaps diluting the message of what is important. The Cincinnati version concentrates on 15 dimensions for rating, though the rubric structure allows distinguishing 32 aspects of teaching. The FAS Continuum can also distinguish 32 dimensions, but some of them seem to overlap and some 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 32 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 National Board assessment 10. These assessments have a narrower focus on specific career stages (novice teachers, accomplished teachers) and emphasize the content needed to certify.

All of the systems have multi-level rubrics or rating scales with more or less specific examples of behaviors that help to define the levels. Rubrics are important for both assessment reliability and validity and for use to help teachers improve instruction. They provide guidance to assessors on what to look for, and what constitutes evidence for performance at each level. When providing feedback to teachers, rubrics communicate the specifics of the vision of instruction the district wants implemented, facilitate assessors in pointing to specific behaviors on which ratings were based, and provide teachers with 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 really behaviorally-based in the same way the others are. The scoring guide describes the characteristics of the evidence as much as or more than referring to teacher behavior. The levels are described based on the characteristics of the evidence rather than on differences in teacher behavior. The Level 4 descriptors define the competencies of accomplished teaching, and the lower levels descriptors generally define what a less convincing response would look like. The basic idea is that a Level 4 response is defined as one that provides clear, consistent, and convincing evidence, Level 3 as providing clear evidence, Level 2 as providing limited evidence, and Level 1 as no evidence that the teacher has a set of competencies. This is likely due to the purpose of the assessment as an indicator of accomplished teaching. The scoring system thus does not have to define lower levels of teaching, but only the degree of evidence that teaching is accomplished. But this does not describe a developmental sequence of teaching competence like the other rubrics.

A third important similarity is that almost all the approaches pay close attention to assessor reliability by requiring a substantial amount of assessor training, and by including provisions for assessor accountability. All of the systems recommend substantial multi-day training in the use of the assessment tools, including training on the process and the content of the rubrics. Four of the systems (Cincinnati, CLASS, PRAXIS III, and TAP) require assessors to demonstrate agreement with assessments of standard vignettes made by master assessors before they are allowed to assess. NIET 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 another assessor to allow examination of assessor calibration. Assessors who deviate from the standards are retrained until they match a criterion or are not allowed to assess.

Some of the approaches have additional features that are 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 the 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 teachers at other points at which consequential decisions will be made based on results. For some uses, multiple assessors are used, typically one of these specialists and a school administrator. The TAP system also uses multiple assessors, though all would be from within the same school. Mentor or master teachers share the assessment responsibility with school administrators, and 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 it was not designed for consequential uses. 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.

Another feature that should improve reliability and validity that is used by some of the observation-based assessment approaches is the requirement for making multiple observations. 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 provides a representative basis to make an assessment, especially for consequences. TAP requires that 4-6 observations be made each year, with at least half of these unannounced. Cincinnati has required 5-6, and now requires 4 for a comprehensive evaluation such as for consequential uses, with three of these unannounced. CLASS does not specify a number of observations, but the procedural handbook does recommend a minimum of a two 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 as few as 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. They are not, however, simply observational systems substitute videos for live observations. The videos are intended to show specific aspects of teaching rather than represent a random sample of teacher behavior. The videos are not the main source of evidence in either system. Both also rely heavily on what teachers write about their practice, guided by prompts. The videos illustrate and confirm what the teacher is writing about. These prompts focus teachers' responses in ways that are intended to explicitly exhibit content knowledge and content related pedagogy. The National Board assessment goes even farther by including six written exercises that are essentially tests of content and pedagogical content knowledge. This is much more efficient than making several live observations in the hope of seeing specific skills being 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.

While 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 have 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 assessors be knowledgeable about the content area. The Cincinnati approach has attempted to match assessors with teachers based on grade and content specialty, but the others do not appear to address this.

Evidence for Reliability and Validity

For use in making consequential human capital management decisions, we want assessment results that are reliable and valid. This section reviews the evidence for the reliability and

validity of assessments made using each of the systems, and then attempts to identify the characteristics of the systems with higher levels of reliability and validity evidence.

Reliability is often considered a basic requirement for consequential uses of assessment results. The aspect of reliability typically of most concern with respect to teaching performance assessments is inter-rater reliability, often represented by inter-rater agreement. There is evidence on inter-rater agreement for five of the assessments: CLASS (Pianta, LaParo, & Hamre, 2008), Cincinnati (Milanowski and White, 2001; Heneman and Milanowski, 2003), NBPTS (National Research Council, 2008), PACT (Pecheone and Chung, 2007), and TAP (Schacter and Thum, 2004). In general, inter-rater agreement is adequate to excellent, with PACT, the NBPTS assessments, 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 HCM decisions with important consequences for teachers.

There are notable commonalities across these systems that we believe contribute to assessment score reliability. First, all have used trained assessors and all have some way to hold assessors accountable for applying the rubric correctly. Second, the rubrics have been developed to be specific, with well-defined distinctions among levels. It is also interesting to consider why CLASS, PACT, and NBPTS scores have especially high reliability. The PACT and MBPTS assessments involve review of standardized artifacts away from the classroom. Teachers are given considerable guidance as to what they are to prepare and submit. Thus there is less "noise" in what is assessed than might be typical in a classroom observation. Assessors also do their scoring apart from the distractions of a normal school day. Beyond the substantial amount of rater training and the careful design of the rubrics, the standardization of observation procedures likely contributes to the high reliability shown by CLASS assessments. It should also be noted that the CLASS reliability studies used researchers as raters rather than school administrators. This was the case with the TAP reliability studies as well. The Cincinnati system also has 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 assessors from outside the school are used.

Assessment validity is a more complex matter to judge. There are a number of types of evidence of validity. One is content validity evidence: evidence that the content of the assessment (what it measures) matches the content of the job. This evidence typically involves expert judgments on how well the assessment method covers or represents the content of the job. Extensive content validity evidence of this type exists for PRAXIS III (Dwyer, 1994; Reynolds, 1995). This content validity evidence also applies to the basic Framework for Teaching, which was built around some of the same competencies as PRAXIS III. Danielson (1996, 2007) connected this evidence to the Framework. There is no independent content validity evidence for the Cincinnati version. In the case of the National Board assessments, substantial content validity evidence also exits, but in this case such evidence was designed to show that the content of the assessment matched the content of the Board's standards for teaching rather than the teaching job or role as such (see Jaeger, 1998; and Moss, 2008). Content validity studies were also done for PACT, showing that this assessment both represented performance dimensions important to teaching and alignment to the California Teaching Performance Expectations (Pecheone and Chung, 2007). The FAS Continuum was derived from the California teaching standards, which

in turn were the subject of a content validation study (Whittaker, Synder, and Freeman, 2001). 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 underway. No content validity evidence was located pertaining to TAP.

Review of this evidence suggests that PRAXIS III and arguably the Framework for Teaching adequately reflect the most important aspects of the teaching role, though in its attempt to be comprehensive, the Framework 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 there assessments' results believes that these teaching standards reflect the kind of teaching they want to encourage. For use in a strategic human capital management system, a district or state would likely want to do its own content validity study to assure that the standards and rubrics of whatever system it chose or developed matched its vision of teaching and the instructional strategies for improving student achievement it has chosen. It may be that content needs to be added, and the Cincinnati and TAP systems are examples of making such additions. The analyses described in the section on competency model content should be of use in starting to think about how these assessments might capture the instructional vision.

Another form of validity evidence of prime interest to most districts or states is the relationship between teaching scores or ratings made using an assessment process and the average achievement of the students taught. This type of evidence is sometimes called criterion-related validity, to reflect the idea that there may be an external standard of performance (the criterion) that we believe the assessment scores should correlate with or predict. For measures of teaching, the currently-accepted criterion is the average value-added achievement in the classrooms in which the teaching is being measured. Such evidence is available for four of the assessments in our review: CLASS, the Framework for Teaching, the National Board assessments, and TAP.

The most research has been done on the National Board's assessments. According to a summary by the National Research Council, average value-added was consistently higher in the classrooms of certified applicants compared to applicants that failed the assessment. Comparing certified teachers to non-applicants, the results are mostly in favor of certified teachers but not as consistently (National Research Council, 2008). Members of our research group have obtained such evidence for three adaptations of the Framework for Teaching, including that used by Cincinnati. We found that on the average teachers with higher total ratings had classrooms with higher average student achievement in reading/language arts and mathematics (Milanowski, Kimball, and Odden, 2005, Gallagher, 2004). Research by Schacter and Thum, (2004) on NIET's TAP evaluation model also found that the average ratings of teachers 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. There has been a fairly substantial amount of research on the pre-K version of CLASS (see Pianta, LaParo, & Hamre, 2008). This research has found consistent positive associations between CLASS measures and student outcomes at the pre-K level. 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, & Morrisonet, 2008). Further research using the K-3 and 5-12 versions is just getting underway. Research has also been conducted on the ancestor of 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). Currently, similar research is underway on PACT. No research of this type using PRAXIS III or the FAS Continuum has yet been located. More details about the research on the relationship between assessment scores and student achievement can be found in Appendix 2.

Our reading of the research is that teaching assessment systems can produce scores with useable levels of criterion-related validity[4]. While 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 is needed. It is notable that the CLASS, Cincinnati, National Board, 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 outside raters, which may 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 HCM 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.

Toward a Specification for Teaching Assessment

Based on our review of the competencies covered by the eight approaches, their assessment procedures, and the evidence for the reliability and validity of assessment scores, this section proposes 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. The specifications attempt

[4] 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, the alignment of the curriculum taught to the tests, and misalignment between the pre and post tests used in the value-added analyses all tend to attenuate any positive relationship. Also, note that strictly speaking all reliability and validity evidence pertains to specific scores, not to the systems themselves. One reason for this is that the scores are typically dependent on how the assessment system was implemented. If assessors are poorly trained, don't observe carefully, or don't follow scoring directions, reliability and validity results are not likely to match these found in the research. To achieve comparable score reliability or validity, the process must be followed as designed.

to present what a state of the art teaching assessment system would look like, based on 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, if the assessment system is not implemented as intended it is unlikely to realize the desired benefits. It is clear from the research on performance evaluation that implementation poses a challenge for organizations in both education (Halverson, Kelley, and Kimball, 2004; Heneman and Milanowski, 2003; Davis, Pool, and Mits-Cash, 2000) and other sectors (Bretz, Milkovich, and Read, 1992; Roberts, 1995). Therefore districts need to specify the details of implementation and develop a plan to carry it out, 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 HCM uses of teaching assessment. High stakes uses (e.g., career ladders or knowledge and skill 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 looking for depth of content knowledge (beyond whether teachers make content errors), asking teachers to demonstrate specific knowledge via a performance task is likely to be more efficient than making multiple observations. To do a good job of monitoring the implementation of the instructional strategy, recognizing good work, and keeping the focus on important performance goals, many short, unannounced observations ("walk throughs") might be more useful than a few full period formal observations

Third, while there probably should be multiple methods of data collection customized to specific HCM uses, all of the methods should be 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 unless the same competencies underlie different methods and uses, the messages received are that the district is at best not serious about any one set of competencies and at worst confused about what it is doing.

Potential Specifications

With these cautions in mind, we outline eight specifications for a state-of-the art district teaching assessment system below. 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 structure of the rubrics need to be applicable to all grade levels, career levels, and subjects. But this should be coupled with customized language in performance dimension definitions and rubrics, when needed, to ensure grade or subject specific instructional strategies or skills are included. The PACT assessments

are a good example of how a single set of basic assessment 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 consequences, the assessment system should include standardized performance tasks that ask teachers to demonstrate this knowledge. It would require too many observations to get a good sample of teacher behaviors indicating the depth of this knowledge. Even if several are done, there is no guarantee that a representative range of content-related knowledge will be observed in use. 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 is done for consequential or summative purposes, the system should include features that promote reliable and valid measurement. These include:

   - multi-level, behaviorally-anchored rating scales or rubrics
   - assessor training
   - an assessor qualification process, such as requiring assessors' ratings of a sample of benchmark cases to match those of an expert panel before allowing them to assess
   - holding assessors accountable for following the process
   - use of multiple assessors or checks of assessment scoring by additional assessors, at least for a sample of decisions
   - 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 should not depend solely on teachers' direct supervisors (e.g., the principal) as assessors. Serious consideration should be given to using assessors from outside the school when assessment results are to be used for major decisions like tenure or career ladder progression.

6. The system should include features that promote teacher learning:

   - assessment results should include enough detail for teachers to understand why they received the scores they did; specific feedback should be provided
   - someone should be trained and responsible for providing coaching and assistance to teachers who want to improve their assessed performance
   - for use in induction and intensive professional development, the assessment should be embedded in a planned set of developmental activities

7. The assessment process should be standardized and documented.

- evidence gathering, evidence interpretation, and evidence evaluation procedures should be spelled out so that they can be implemented uniformly
- assessment conditions such as number and length of observations, length of video clips, prompts for performance tasks should be specified

8. The system should maximize the use of technology to minimize workload and improve administrative feasibility.

- the use of web-based data collection and scoring and video for recording behavior should be designed into the system
- technology-enabled mentoring such as that available to users of CLASS (Hadden & Pianta, 2006) should also be included to help provide high quality professional development directly related to the competencies being assessed

A Preliminary Roadmap for Designing a Teaching Assessment for HCM

The development of a teaching assessment for use in a strategic human capital management system is a complex undertaking. While this paper does not presume to offer a complete description of how this might be done, we do think the process outlined below can help districts or states considering such an assessment think about how to get started and what resources would be 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.

Process Steps

1. Develop a competency model on which to base the assessment system and other HCM programs.

This process would begin with a review the district or state vision of instruction and strategies for improving student achievement. Bring together a group of people knowledgeable about the vision, strategies, and how they would be implemented to develop a list of what do teachers need to know and be able to do to carry out the vision and strategies. The group would also review existing competency models, including those underlying the assessments we have reviewed, the current state teaching standards, or 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 would develop a competency model that reflects the most important aspects of the vision and drive successful implementation of the strategies. This model should be concise, avoiding including all desirable competencies and focus on the most critical ones. 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 model. While most existing systems are likely to require some modifications to fit individual state or district needs, in particular to reflect its specific performance improvement strategies, there will likely be a lot of basic content that can be taken over from an existing competency model and assessment approach. Developing a statewide or multi-state model, then customizing it to the needs of individual districts seems a plausible and economical approach. It is likely that core (common across districts and schools) and non-core (specific to districts and schools) performance competencies can be distinguished. It would be possible to develop a system to assess core competencies that could be used by multiple districts, then add non-core content or supplementary assessment procedures. 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 customization of the Framework for Teaching by Cincinnati is another type of example. Danielson designed the Framework to be generic and comprehensive, but advised customization. Cincinnati followed this advice, making fairly substantial changes to reduce the number of dimensions and concentrate teaching practices considered the most important to improving achievement. While Cincinnati might have been able to develop their system from scratch, beginning with the Framework jumpstarted the development process and avoided reinventing the wheel.

2. Decide on a high-leverage initial use of the assessment for the initial development effort. This would likely be an assessment that would be tied to some important HCM decision such as initial selection, professional licensure, granting tenure, or career ladder movement.

3. Develop the assessment for the initial use.

To begin this process, it is advisable to develop an assessment plan. Such a plan specifies the assessment uses, the competencies to be assessed, and the methods of collecting evidence on and rating the level of the competencies. Different competencies are best assessed using different forms of evidence collection, such as performance tasks, videos, artifacts like lesson plans and student work, and live observations. It is also important to consider the efficiency of different methods and their potential reliability and validity under operational conditions. Based on the potential use of an assessment for a tenure decision, we summarize some of the considerations for an assessment plan below.

A critical requirement for such a high leverage use is high reliability and validity. This suggests that where possible the assessment should be externally scored by trained assessors with subject/grade level expertise. An assessment based on a set performance tasks such as the NBPTS or PACT assessments seems like a good candidate for measuring as many of the major competencies as possible.

One approach to 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, provide sample materials, assignments, and assessments, and describe how instruction

would be differentiated for a high performing, average performing, and struggling student. A video could be included showing how the teacher introduced the unit, and one showing how s/he taught a key concept. Teachers would also be asked to explain their decisions and reflect on the success of the unit. In order 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 cannot 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. This suggests classroom observations for these dimensions, using an observational rubric such as CLASS or an adaptation of the relevant parts of the Framework.

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. In this pilot, the degree of fidelity of implementation to the assessment procedures would be studied, as would be the feasibility of administration and workload imposed on administrators and participants. Measurement properties such as reliability and discrimination would also be assessed. Teacher, administrator, and assessor reactions should also be studied. This information would surface most implementation problems. Changes to the assessment design would then be made before use for consequential decisions.

5. Analyze assessment needs for other HCM uses and develop supporting assessments.

We strongly recommend that the assessment system be incorporated into, and aligned with, the district's human resource management system. This will make maximum use of 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 HCM uses. For example, performance tasks such as used by PACT and the NBPTS do not measure of typical behavior, so such an assessment is 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 HCM uses, and choose a data collection system that best fits those uses. Table 4 presents some suggestions. Note that the key to developing an aligned system is to design each assessment process beginning from the competency model. While some of the assessments might focus on different subsets of the competencies, 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 a third set. Often these 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 of teachers and school leaders about what the district values.

<Insert Table 4 Here>

It may be useful for a district to consider analyzing the alignment of the entire human resource management system as part of the process of aligning assessments. This 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 and professional development programs that do not support teachers in acquiring the competencies they need to implement the competencies assessed for tenure or career progression. The human resources alignment analysis process developed by Heneman and Milanowski (2004, 2009) may be a useful model for a district alignment analysis.

6. Collect reliability and validity evidence from post-pilot administration.

During the first full administration and periodically afterward, system designers should collect information on reliability of scores, such as inter-assessor agreement. Reliability is not only an important condition for valid measurement, but it is also important to establishing the credibility of the system with teachers. Prior experiences with teaching practice assessment, such as performance evaluations, may have convinced many of them that practice ratings depend on who is doing the observing as much as the teaching observed. Collecting and reporting reliability evidence may help to counteract some of the skepticism teachers may have 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. While it is unrealistic to expect a strong relationship[5], 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 to .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 that the measurement procedures actually being used are not reliable or being implemented as intended. Calculating these correlations provides districts with important information that can be used to justify the use of teaching assessments for human capital management decisions and 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 that the assessment system does not 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 re-checked (especially if assessments were made using videos and artifacts) to look for leniency. High rated teachers could be

[5] There are many factors that weaken the relationship, including some unreliability in both value added and teaching assessment scores, and the fact that teaching assessment scores and value added are not measuring the same thing. 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 can't control, a very high correlation would actually be suspicious.

interviewed or observed to look for practices that differentiate high versus low value-added classrooms but might not be reflected in the assessment system.

7. Consider combining teaching practice assessment and measures of student outcomes such as value-added estimates of classroom productivity.

At the beginning of this paper, we argued that measuring teaching using only estimates of value-added student achievement was not sufficient to improve instruction nor sufficiently valid for all HCM uses. However, outcome measures such as value-added have important roles to play in teaching assessment. The persistent problem of leniency in evaluations of practice based on administrator judgment (see Weisberg et al., 2009 for the latest documentation) suggests that these judgments are also not sufficient. Besides using value-added or similar outcome measures to provide validity evidence for teaching assessment scores, the two approaches can work together in two other ways: calibration and complementary measurement.

Calibration refers to the use of value-added estimates to examine how well specific assessors or sets of assessors might be doing and improve future judgments. Looking at the relationship between assessment scores and value-added estimates for assessors or assessor groups can identify differences that could indicate that some assessors are having trouble applying the process. For example, an assessor may rate most teachers at the highest level, thought these 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 she or he was applying the system, and if problems were apparent, could be provided with additional training. Such interviews can also provide indicators of how well assessors are following the process and their motivation to make accurate judgments (Kimball & Milanowski, 2009). Even simply showing assessors graphs of the relationship between value-added 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 practice assessments based on judgment for some important human capital management decisions. Using both would provide multiple indicators of teaching performance, recognizing the importance of outcomes and the importance of 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 three years of positive value-added estimates (which would show the teacher's classroom was consistently achieving above-average learning gains). Use of both measures would accomplish three things. First, it would guard against lenient practice assessment allowing teachers who are just going through the motions without contributing substantially to student learning from achieving tenure or a higher career level. Second, it sends the message that both teaching practice and results are important. Third, it recognizes that both measures have error, and to the extent the two are correlated, the combination of the two will have less error than each by itself.

As mentioned in the introduction, there are some legitimate concerns about the validity of using a value-added estimate of classroom productivity as an indicator of teacher performance. The prescription for addressing these concerns has typically been to use estimates based on multiple years (e.g., Koedel and Betts, 2009), typically three. Taking this advice, there are two career points at which it would seem logical to combine value-added with a teaching practice assessment. The first is the tenure decision. One could require a proficient level of practice, a specific level of value-added, and a recommendation from the appropriate school leaders for tenure. The second is movement to the highest level of a career ladder or knowledge and skill-based pay schedule. Here it would seem logical to require the teacher to exhibit both exemplary practice and above-average value-added.

An issue that would have to be addressed is the minimum value-added score required. Since value-added estimates are relative to the other teachers in the state or district teaching force, there is no natural cut-off point that represents acceptable performance for tenure. While it may seem attractive to require a teacher to produce the average value-added, there are drawbacks to this. Since somewhere near half of the teachers will of necessity be below the average, a state or district may not be able to dismiss nearly half of its new teachers. There may be insufficient supply, and this also requires additional resources directed to recruitment, selection, and induction. There is need for additional research on this point. One 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, then require value-added levels sufficient to maintain the trajectory.

A second issue is what to require for teachers of non-tested subjects. While more test development is always an option, in the short run it may be easier to adapt goal-setting approaches such as those used in Denver's ProComp and in Orange County, Florida. 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 combination of value-added with judgmental practice assessment in the ways discussed above seems like a fruitful partnership. It could produce both 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.

References

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.

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.

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.

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* . NY: MacMillan, 799-

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. (2008). Is it just a bad class? Assessing the stability of measured teacher performance. Center on Reinventing Public Education Working Paper 2008-5. Center on Reinventing Public Education, Seattle, WA.

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.), *Research and Theory in Educational Administration,* Volume 3. Greenwich, CT: Information Age Publishing, .

Heneman, H.G. III, Milanowski, A., Kimball, S.M., and Odden, A. (2006). Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay. CPRE Policy Brief RB-45. Philadelphia, PA. Consortium for Policy Research In Education.

Heneman, H.G. III, and 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, and Milanowski, A.T. (2009). Analyzing human resource practice alignment. Paper prepared for the March 2009 SMHC District Network meeting. Available at: http://www.smhc-cpre.org/resources/district-reform-network/hr-alignment-analysis/

Heneman, H.G. III, and Milanowski, A.T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education,* 17:3, 171-195.

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.

Kimball, S. M., and 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.

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

Levy, P.E., and 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.

McCafferey, D.F., Sass, T.R., and Lockwood, J.R. ((2008). The intertemporal stability of teacher effect estimates. National Center on Performance Incentives. Nashville, TN. Available at: http://www.performanceincentives.org/data/files/news/PapersNews/McCaffrey_et_al_2008.pdf

Milanowski, A.T., Kimball, S.M., and Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A.E. Schwartz, R. Rubenstein, and J. Zabel (eds.) *Measuring School Performance and Efficiency: Implications for Practice and Research,* the 2005 American Educational Finance Association Yearbook,137-159.

Milanowski, A.T., and White, B. (2001). Rating agreement in the 2000-2001 implementation of the Concinnati Public Schools/Cincinnati Federation of Teachers teacher evaluation system. Madison, WI, Unpublished.

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 and J. Hattie (eds.) *Assessing Teachers for Professional Certification: The First Decade of the National Board for Professional Teaching Standards*. Oxford, UK: Elsevier, 257-312.

Murphy, K.R., and 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.* Committee on Evaluation of Teacher Certification by the National Board for Professional Teaching Standards, Milton D. Hakel, J.A. Koenig, and S.W. Elliott,

Eds. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

Pecheone, R.L., and Chung, R.R. (2007). Technical report of the Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003-04 pliot year. Palo Alto, CA: PACT Consortium. Available at: 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., and Hamre, B.K. (2008). *Classroom Assessment Scoring System Manual K-3*. Baltimore, MD.: Paul H. Brookes Publishing.

Prince, C.D., Schuermann, P.J., Guthrie, J.W., Witham, P.J., Milanowski, A. T., and Thorn, C.A. (2008) The Other 69%: Fairly Rewarding the Performance of Teachers in Non-tested Subjects and Grades. Center for Educator Compensation Reform. U.S. Dept. of Education, Office of Elementary and Secondary Education, Washington, DC. Available at 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.

Schacter, J., and Thum, Y. (2004). Paying for High- and Low-Quality Teaching. *Economics of Education Review*, *23*(4), 411-430.

Weisberg, D., Sexton, S., Mulhern, J., and Keeling, D. (2009). The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. The New Teacher Project. Available at: http://widgeteffect.org/downloads/TheWidgetEffect.pdf

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

Wilson, M, Hallam, P.J., Pecheone, R., and 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. Submitted to Educational Evaluation and Policy Analysis.

Table 1
Overview of Teaching Assessment Systems Reviewed

System	Original Purpose	Extent of Use	Theoretical Perspective	Instrument Structure	Specialization
CLASS K-3	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 was used for research purposes & in Head Start. K-3 and middle/secondary versions are now being used in several districts. Also used in U of VA's electronic teacher professional development program and by ABCTE for Distinguished Teacher certification.	Child developmental theory & research showing importance of quality of teacher-student interactions, including social and emotional functioning in the classroom.	Ten dimensions grouped within three domains. Rubrics define 3 levels using behavioral descriptions, but allow scoring within levels so that scores can range from 1-7	The 3 CLASS versions span pre-K to high school, and apply to all experience levels & content areas. The version included in this study applies to K-3.
FAS-CTD	New teacher induction & mentoring	Forty-six induction programs in California; also used in NYC and Chicago induction programs.	Holistic view of teaching emphasizing content, student diversity & inclusion, differentiation of instruction, student engagement, & self-directed learning	Six standards (domains) with 5-6 components per standard; 32 total rateable dimensions. Rubrics define 5 levels all with behavioral descriptions	Focus on novice teachers at all grade levels & content areas; but higher rubric levels fit experienced teachers as well.
FFT- Original	Formative tool for promoting conversations about good teaching. Suggested uses included self-assessment, induction & mentoring, peer coaching, and clinical supervision	No data on extent of use available, but anecdotal evidence suggests use in some form at least 200 districts of all sizes & types.	Intended to be a comprehensive representation of generic teaching activities applicable to almost all K-12 settings. Emphasizes aspects of constructivism.	Twenty-two components grouped into 4 domains; components are further divided into elements. There are 66 elements in all. Rubrics define 4 levels using behavioral descriptions.	Intended to apply to all career & grade levels, and content areas.
FFT-Cincinnati	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 standards grouped into 4 domains. Rubrics define 4 levels with behavioral descriptions.	Intended to apply to all career & grade levels, and content areas.

Table 1, continued

System	Original Purpose	Extent of Use	Theoretical Perspective	Instrument Structure	Specialization
NBPTS-EA Mathematics	To assess accomplished teaching practice as part of a voluntary certification system intended to recognize high quality teachers	National Board certification is supported, recognized, or rewarded in all 50 states and hundreds of districts. There are about 74,000 certified teachers.	Based on Board's 5 core propositions: teachers 1) are committed to students and their learning; 2) know the subjects they teach and how to teach those subjects; 3) are responsible for managing and monitoring student learning; 4) think systematically about their practices and learn from experience; 5) are members of learning communities.	Four portfolio entries developed by teacher and 6 assessment center exercises requiring teacher constructed response. Rubrics define 4 levels with specific anchors describing characteristics of response at each level.	Experienced teachers (3+ years of experience) in middle school and early high school mathematics. Similar assessments used for 24 other certification areas.
PACT - Mathematics	New teacher initial licensure	Thirty-two California university and district teacher preparation programs	"Authentic" assessment in place of paper and pencil tests; based on a plan, instruct, assess, and reflect cycle, with special attention to subject-specific pedagogy and the teaching of English language learners.	Twelve guiding questions divided among 5 domains are scored. Rubrics define 4 levels with specific anchors describing characteristics of response at each level.	New teachers to be licensed to teach middle & high school mathematics. Similar PACT assessments used for 25 other areas of licensure, including elementary grade generalist.
PRAXIS III	Teacher licensure	Used in Ohio & Arkansas for licensure.	Development was guided by assumptions that effective teaching requires both action and decision making, learning is a process of the active construction of knowledge, and since good teaching depends on the subject there is no teaching style that is best for all contexts.	Nineteen dimensions grouped into 4 domains. Rubrics define 3 levels with specific anchors; two intermediate levels can be scored between the anchored levels providing 5 possible ratings.	New teachers in all content areas and grade levels.

Table 1, continued

System	Original Purpose	Extent of Use	Theoretical Perspective	Instrument Structure	Specialization
NIET- TAP	Formative and summative evaluation of teachers in TAP schools, including use in career ladder program & possible use for pay bonuses.	TAP website cites use in 220 schools as of Fall 2008. Used in many of the Teacher Incentive Fund sites, some Q-Comp sites in Minnesota, and several districts in Louisiana, North and South Carolina, and Texas.	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. For 3 domains, rubrics define 3 levels with behavioral anchors. Raters can score in between levels providing 5 possible ratings. Scales for Responsibilities dimensions are simply sentences that are rated 1-5 with frequency anchors at 5, 3, and 1 levels. Local users customize this part of system.	Intended to apply to all career & grade levels, and content areas.

Table 2
Coverage* of Eight Important Teaching Competencies

System	Attention to Student Standards	Use of Formative Assessment to Guide	Differentiation of Instruction	Student Engagement	Higher Order Thinking Skills	Content Knowledge and Pedagogical Content Knowledge	Personalized Relationships with Students	High Expectations
CLASS K-3	0 0 0	0 0 0	2 20% 1	3 30% 1	3 30% 1	0 0 0	2 20% 1	1 10% 0
Framework for Teaching-Cincinnati	3 20% 1	1 7% 0	2 13% 0	3 20% 0	2 13% 1	3 20% 1	1 7% 1	1 7% 1
Framework for Teaching-Original	0 0 0	1 5% 0	4 18% 0	5 23% 1	3 14% 0	3 14% 1	1 5% 0	3 14% 1
NTC- FAS Continuum	2 6% 0	2 6% 2	5 16% 2	5 16% 4	4 13% 2	2 6% 2	1 3% 0	2 6% 0
NBPTS- EA Mathematics	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-Mathematics	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
NIET-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

*The top number in each box is the number of performance dimensions with rubrics that refer to the competency; middle number is the percent of these dimensions that refer to the competency; 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

Aspect of Teaching	PRAXIS III	FFT	Cincinnati	TAP	FAS CTD	CLASS	PACT-Math	NBPTS- EA Math
Knowledge of Students	■	■	■	■	□		○	○
Appropriate Instructional Goals	■	■	○	■	○		○	○
Communicating Instructional Goals to Students	□		□	□	■	□		
Assessment Aligned to Goals	■	■	■	□			○	
Planning Coherent Instruction	□	■	□	■	■		□	○
Multiple Assessment Methods	□		□	□	■		□	
Assessment Used to Plan/Adjust Instruction	□	■	○	○	■		■	○
Positive Relationships with Students	■	■	■	□	■	■		
Fair, Inclusive Learning Environment	■				■			○
Managing Classroom Procedures	○	■	■	○	■	■		
Maximizing Use of Instructional Time	□	□	□	□	■	■		
Managing Student Behavior	■	■	■	■	■	■	○	
Student Responsibility for Classroom Behavior		□			□	○		
Physical Organization of Classroom	■	■		○	■	○		

■ = 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 3, continued
Additional Common Teaching Competencies Found in Three or More Assessment Approaches

Aspect of Teaching	PRAXIS III	FFT	Cincinnati	TAP	FAS CTD	CLASS	PACT-Math	NBPTS-EA Math
Questioning/Discussion Techniques		■	■	■	□	□		○
Quality of Feedback to Students	○	■	■	■	○	■	■	○
Use Variety of Instructional Strategies	○		□		■	□		
Building on Student Experiences/Interests		□	□	○	■	□	□	○
Grouping of Students		■	○	■	□			○
Student Initiative in Learning		□	○	○	■	□		○
Adaptation of Lesson/Plan		■	□	○	■	□	■	○
Reflection on Practice	■	■	■	■	■		■	□
Communication with Families	■	■	■		■			○
Cooperation with Colleagues	■	■	■		■			○
Pursuit of Professional Development		■	■	■	□			○

■ = 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
Suggested Data Collection and Assessment Methods for Human Capital Management Uses

HCM Use	Suggested Data Collection & 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 & Mentoring	Observation tool focusing on competencies being developed by induction & mentoring program, including pre- and post-observation conferences.
Professional Development	"Work samples" using videos of lessons and artifacts around an instructional unit that can be assessed off site by expert who would also provide specific feedback & coaching. Could also be used by teachers to prepare for consequential assessments or by lesson study groups.
Performance Management	a) Teacher evaluation tool based on multiple live observations by school leaders. Cycle of observations would include pre-arranged observations so that assessor could see how well teacher was implementing suggestions. b) Walk through tools focusing of simple judgments by school leaders of whether 1-3 readily-observable competencies are being displayed.
Compensation (such as a career ladder or knowledge & skill-based pay system)	Performance assessment based on instructional units scored by external assessors. Might require at least a "proficient" performance evaluation rating by school administrators to be eligible for movement. [6] Could also require evidence of student learning for progression to highest levels.

[6] Additional considerations on using teaching assessments as part of a teacher compensation system can be found in the CPRE Policy Brief "Standards-Based Teacher Evaluation as a Foundation for Knowledge- and Skill-Based Pay" (Heneman, Milanowski, Kimball, & Odden, 2006).