CHAPTER EIGHT

Approaches to Annual Performance Assessment

Mary M. Kennedy

Michigan State University

There used to be a popular expression about not knowing anything about art but knowing what you liked. The implication was that one needed no specialized knowledge to appreciate art. A similar claim could be made about teaching. Few people believe they need specialized training to tell good teaching from bad, and virtually everyone has strong feelings about the quality of teaching they encounter as they or their children move through the school system. Everyone can recognize good teaching when they see it.

Everyone, that is, except school administrators, who have been struggling for an entire century to devise strategies for defining and measuring teaching practice and for discriminating between good teaching and bad. Throughout this century, they have tinkered with a variety of checklists, forms, rating scales, and measurements of all sorts and have yet to achieve a consensus on either definition or procedure. Even the language we have used to describe good teaching has changed over time. At the beginning of the twentieth century, when industrial efficiency was a new and fashionable concept, school administrators sought teachers who were efficient. Then they sought teachers who were virtuous. Then teachers who were not neurotic. Later in the century, they sought teachers who had specific competencies, and toward the end of it, they sought professionalism and expertise. Today we tend to want teachers who meet professional standards.
These changes in terminology illustrate the variety of ways people have thought about teaching over the century. But they reflect changes in fashion more than changes in the underlying concepts. For example, teachers evaluation forms ostensibly assessing efficiency often included virtues, and those assessing competencies included some items that look like contemporary standards. So while the century has seen many changes in terminology and methods of documentation, it has nearly always incorporated into these terms a wide variety of competing, and sometimes even contradictory, criteria for teaching quality.

We face two central problems. One is that we actually hold numerous public values, or criteria for defining teaching quality, and sometimes they contradict one another. As a society, we want our youngsters to learn particular content, but we also want them to be nurtured, to be developed into good citizens, and to be motivated to participate productively in society. We want teachers to be role models for moral and ethical behavior and to create positive climates for learning in their classrooms, but we also want them to be efficient and goal oriented. We want teachers to treat all children equally, but we also want them to respond to each child’s unique needs. We want them to be caring and nurturing, but we also want them to be rigorous and demanding. We want them to cultivate cooperation in students, yet enable them to compete in later life. These different ideas wax and wane in their social popularity, and they strain the education system. When they get translated into formal assessment instruments, it should be no surprise that they lack coherence or consistency. This is the problem of accommodating public values.

The second problem is that as an enterprise, teaching events have meanings that are difficult to capture on a form. No matter which public value we choose to evaluate, we quickly discover that we cannot directly see it. It must be inferred from sequences and patterns of events. Gergen (1982) gives a useful example of this problem. He invites us to imagine ourselves at a party, and across the room, we see Ross touch Laura’s hair. Gergen asks what action we have actually observed. We are not illuminated to know that Ross’s hand moved at a particular velocity along a particular trajectory. We need to know what the action means, what its significance is. If Ross recently said he loved Laura, we might infer that his touch was an act of affection. If Ross and Laura had recently had an argument, perhaps the gesture was an act of derision. If they had just met, perhaps it was flirtatious. Or perhaps it was merely an effort to flick off a piece of lint. Gergen’s point is that meaning derives from our knowledge of history and context. If we know these other things, we can interpret the gesture. And we continue to reconstruct the meaning of events as we gather more information. Does Laura brush his hand away or smile warmly at him? Each new piece of information we obtain about these two people adds new layers of meaning to this brief action.
Meaning by itself is not difficult to infer, but it is very difficult to infer if we want an objective evaluation, one that multiple observers would agree on. The problem is that while two observers could agree that a touch occurred, thus giving us interrater agreement, they would likely infer different meanings from the event, especially if they were privy to different subsets of historical or contextual information. The problem we face when evaluating teaching practices is that we need, on one side, to capture the meaning of events, yet at the same time, we need to ensure that our instruments meet standards of fairness, reliability, and objectivity. This is the problem of balancing meaning and objectivity. We seek an approach to evaluation that allows us to evaluate the meaning of events, but to do so reasonably objectively.

The combination of these two problems has made teacher assessment an especially perplexing problem. My aim in this chapter is to review our century-long history of efforts to solve these two problems: the problem of accommodating public values and the problem of balancing attention to objectivity with attention to meaning. This history reveals a persistent, century-long tendency to overlook the intellectual core of teaching itself. Our efforts to accommodate public values have focused on nearly every aspect of teaching except its intellectual purpose, and our efforts to generate reliable and objective instruments have similarly overlooked the intellectual work of teaching and learning.

This is not to say that we have made no progress, for we have. Our efforts to accommodate public values have progressed from an early focus on the personal qualities of teachers as individuals—their dress, friendliness, virtue, and so forth—to a focus on what they actually do. In regard to the second problem, we have progressed from highly judgmental, high-inference rating scales to our current reliance on performance rubrics. The two shifts, from personal qualities to practice and from rating scales to performance rubrics, are substantial, and they make it possible for us to now look at the intellectual work of teaching.

**ACCOMMODATING PUBLIC VALUES**

During this past century, Americans have embraced several different conceptions of the Good Teacher, and teacher educators have tried to give clarity to these conceptions. One early attempt (Charters & Waples, 1929) came up with these twenty-five essential qualities: Adaptability, Attractiveness, Breadth of Interest, Carefulness, Consideration, Cooperation, Dependability, Enthusiasm, Fluency, Forcefulness, Good Judgment, Health, Honesty, Industry, Leadership, Magnetism, Neatness, Open-Mindedness, Originality, Progressiveness,
Promptness, Refinement, Scholarship, Self-Control, and Thrift. Teachers, in other words, had to do better than Boy Scouts.

This list is revealing for several reasons. One is that it is a list of qualities of teachers, not qualities of teaching. Another is that it is a very long list and represents a remarkable diversity of ideas. In fact, these ideals may be incompatible. Is it possible for one individual to be both dependable and adaptable, for instance, or might adaptability lead to some lapses in dependability? Is it possible to be both enthusiastic and also self-controlled, or might one’s enthusiasm trigger lapses in self-control? Some of these qualities, such as neatness, good judgment, promptness, and self-control, suggest a great interest in propriety, something Waller (1932/1961) commented on in his early analysis of teachers. But other qualities, such as forcefulness, magnetism, and open-mindedness, suggest a personality that may occasionally be unconventional and even controversial. These early documents contain no discussion or even recognition of potential contradictions among valued qualities of teachers. In that sense, they represent public values, which are frequently contradictory, far more than they represent the kind of a rational, internally consistent model of teaching with explicit constructs that can be measured. And in fact, Charters and Waples’s intention was to capture public values. Their list was generated from interviews with parents, teachers, administrators, and teacher educators. It was not intended to be a list of precise terms but rather a list of valued qualities.

A look at the teacher assessments in use during that early period reveals a shorter list of qualities, but also numerous variations on each theme, thus suggesting a lack of precision. Nearly all early evaluations took the form of rating scales—that is, teachers were rated as good, fair, poor, and so forth on a list of qualities. Some qualities consisted of character traits, like those listed by Charters and Waples, while others resided in teaching practice itself. As is the case today, each district created its own assessment form, and they were quite various. Occasionally someone would survey school districts to learn what criteria they were using. One such early survey (Boyce, 1915) found that the most frequently evaluated qualities were these, listed in order of the percentage of rating forms that included them:

37 percent discipline
23 percent instructional skill
23 percent cooperation and loyalty
22 percent scholarship and education (this refers to their educational background, not their personal inclinations)

Although this list looks relatively parsimonious, notice that none of these criteria appeared in even half the rating systems. Most appeared in fewer than
school people routinely went through motions such as interviewing job candidates and evaluating teachers as if they could recognize a good teacher when they saw one, but in fact, they lacked real knowledge about how to recognize a good teacher.

Even by midcentury, the field had failed to achieve consensus. In 1945, Reavis and Cooper examined eighty-five local rating scales and tallied the number of items that fell into each of seven general categories of values. Their tallies looked like this:

- 329 items evaluating social relations (for example, appearance, tact, loyalty, relations with students, leadership, tolerance, courtesy, participation in community activities, cooperativeness)
- 298 items evaluating instructional skill (for example, presenting lessons, remedial instruction, questioning, evaluating, making assignments, preparing lessons)
- 229 items evaluating personal characteristics (for example, emotions, morality, appearance, attitude, judgment, honesty, humor, use of English)
- 217 items evaluating noninstructional school service (for example, clerical work, safeguarding supplies, extracurricular activities, discipline and guidance)
- 168 items evaluating professional qualifications (for example, level of education, philosophy of education, knowledge of subject, professional attitude)
152 items evaluating habits of work (for example, initiative, punctuality, resourcefulness, efficiency, dependability, originality, ability to plan and organize)

146 items evaluating pupil results (for example, pupil initiative, response, participation, habits, spirit, attention, participation in classroom control)

Most of these efforts to survey the field were intended to find common ground rather than to critique current practice. The first survey I found that critically examined the content of these instruments appeared in Wood and Pohland (1979). Their analysis focused not on whole instruments, but instead on the 1,928 discrete items contained in their sample of instruments, and they raised questions about the content they found. First, they noticed that only a relatively small fraction of items had to do specifically with teaching itself; most had to do with personal qualities, class management, citizenship within the school, or other relatively tangential qualities. Second, they pointed out that the particular personal qualities itemized tended to emphasize conformity with the status quo rather than striving for improvement. For example, the rating forms included criteria such as punctuality, dependability, and loyalty rather than criteria such as creativity, leadership, or initiative.

So for more than half of the twentieth century, local efforts to evaluate teachers could be characterized by these important features. First, they incorporated a wide range of valued qualities, most of which were very general qualities and not well defined. Second, they were remarkably various and remarkably idiosyncratic with respect to the particular qualities they included. Third, many, if not most, of the qualities referred to the teachers’ personality or moral character rather than to teaching practices. And finally, the scoring systems relied heavily on subjective judgment. The heavy attention to teachers’ decorum and moral character suggests that teachers were expected to convey proper behavior more than substantive ideas.

BALANCING OBJECTIVITY AND MEANINGFULNESS

The second problem we confront in devising performance assessments is developing evaluation forms that are sufficiently objective that multiple observers would document the same thing when they watched the same teacher and yet sensitive enough to capture meaningful events in the classroom. The problem of objectively recognizing good teaching was also apparent relatively early in the twentieth century, almost as soon as evaluation systems were first developed. As researchers began to examine these local efforts, they found not only that district-based rating forms were quite various in their content (Boyce, 1915; King, 1925), but also that they were quite various in their
ability to predict student achievement gains (Hill, 1921), and they suffered from very low interrater agreements (Barr, 1929). In effect, these rating scales formalized the kind of ad hoc recognizing that parents and others do when they meet or observe teachers. Supervisors believed they could recognize good teachers when they saw them, but different supervisors noticed and valued different things. Their rating sheets simply captured on paper their subjective and impressionistic judgments. In one early study, Barr (1924) asked teacher supervisors what they looked for when evaluating teachers and obtained 131 different qualities from about a hundred supervisors. Significantly, over half of these qualities were mentioned by only one person, suggesting that the ability to recognize good teaching was distinctly idiosyncratic.

Perhaps because early attention was focused on the problem of sorting out values and agreeing on a set of criteria for good teaching, attention to the measurement properties of teacher evaluation instruments was postponed until the second half of the twentieth century, when researchers began to suspect that the problem of teacher assessment lay in the instrumentation and scoring procedures rather than in the criteria themselves. They made a distinction between high-inference instruments, meaning those that required a high degree of subjective judgment on the part of the observer, and low-inference instruments, meaning items that used carefully worded prompts to reduce ambiguity and reduce the need for interpretation, and they began to develop new low-inference techniques for observing teaching. These instruments forced observers to check off or tally discrete behavioral events rather than rating global impressions, thus increasing the likelihood that two observers would agree about what they just saw. Use of these instruments greatly increases the likelihood that two observers would rate a given event in the same way. Exhibits 8.1 and 8.2 illustrate these two approaches to assessment. Each example is a small portion of an evaluation form.

Exhibit 8.1 comes from an early rating system used by the Oakland school system. The criteria themselves are very broad, and the rater is given substantial latitude in judging the teacher on each of these criteria. At that time, it was assumed that the qualities of good teaching were self-evident, so that anyone could fill out such a form without being trained. Anyone could recognize a good teacher when they saw one. Exhibit 8.2 comes from a system used by the state of Florida from the 1980s until the early 2000s. It focuses on very specific events that are expected to occur in any given classroom lesson. Observers have to be trained in the meaning of each coding category and on rules for counting or not counting particular types of events.

The movement toward low-inference measurements occurred in part because classroom instruction had become an object for research, in addition to being an object for administrative oversight. Researchers came to see classroom observation as a tool for learning more about the nature of
Exhibit 8.1  Excerpt from an Early High-Inference Form: Oakland Public Schools, circa 1945

<table>
<thead>
<tr>
<th>Character</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Possesses highest ideals</td>
<td>Possesses good ideals</td>
<td>Possesses low ideals</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personality</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Winning</td>
<td>Agreeable</td>
<td>Unattractive</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Disposition</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Even, cheerful, pleasant, fine sense of humor</td>
<td>Moderate amiability and balance</td>
<td>Erratic, morose, unpleasant</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Personal Appearance</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neat, clean, dressed tastefully and becomingly</td>
<td>Reasonably well dressed</td>
<td>Careless in cleanliness and dress</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mental Alertness</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Outstanding initiative and imagination; keenly alert at all times</td>
<td>Reasonable amount</td>
<td>Passive</td>
<td></td>
</tr>
</tbody>
</table>
### Exhibit 8.2  Excerpt from a Low-Inference Form: Florida Performance Measurement System, circa 1990

<table>
<thead>
<tr>
<th>Domain</th>
<th>Tot Freq</th>
<th>Frequency</th>
<th>Frequency</th>
<th>Tot Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0 Instruction organization and development</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Begins instruction promptly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Handles materials in an orderly manner</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Orient students to class work/maintains academic focus</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Conducts beginning/ending review</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Questions: Academic comprehension/lesson development</td>
<td>a. Single factual</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b. Requires analysis/reasons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>c.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Recognizes response/amplifies/gives correct feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Gives specific academic praise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Provides for practice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Gives directions/assigns/checks comprehension of homework, seat-work assignments/gives feedback</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Circulates and assists students</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 1. Delays
- 2. Does not organize materials systematically
- 3. Allows talk/activity unrelated to subject
- 5a. Allows unison response
- 5b. Poses multiple questions asked as one
- 5c. Poses nonacademic questions/nonacademic procedural questions
- 6. Ignores student or response/expresses sarcasm, disgust, harshness
- 7. Uses general, nonspecific praise
- 8. Extends discourse, changes topic with no practice
- 9. Gives inadequate directions on homework/no feedback
- 10. Remains at desk/circulates inadequately
instruction and the relationship between teaching practices and student responses. They became interested in specific instructional processes and began to abandon interest in teachers' personal qualities. Two themes guided much of this development. First, researchers hoped to learn more about the relationship between classroom practice and student learning, and second, they wanted to devise instruments that had greater precision.

Interest in low-inference measurement shifted attention away from what to measure and toward how to measure it. Presumably better methods could improve both precision and interrater agreement. Analysts evaluated the strategies used to document teaching, the coding systems, training procedures, interrater agreements, and procedural safeguards (for illustrative reviews of these issues, see Dwyer & Stufflebeam, 1996; Evertson & Green, 1986). The third American Educational Research Association (AERA) handbook of research on teaching included, for the first time, a chapter devoted entirely to observation as a form of data collection (Evertson & Green, 1986). It examines every aspect of observation systems, including how units of behavior are defined, how scales are developed, how raters are trained, and how data are stored and aggregated. It does not, however, address the question of what to measure.

But the problem of meaningfulness remained. Researchers still had to make decisions about what to document when they observed classrooms, and they subscribed to many different theories and hypotheses about what was most important in the classroom. A book edited by Biddle and Ellena (1964) illustrates this variety. Each chapter of the volume describes a prominent and substantial program of research on teaching, and the programs are almost completely unrelated to one another. One focused on teacher reasoning (Turner, 1964), another on classroom atmosphere (Ryans, 1964), another on content representations (Meux & Smith, 1964), and another on “restrictive” versus “inviting” comments from teachers (Flanders, 1964). By the late 1960s, research-oriented observation systems had become so various that a multivolume compilation was produced (Simon & Boyer, 1967a, 1967b, 1970, 1974) to catalogue all the available instruments. In a major literature review appearing around that time, Rosenshine and Furst (1973) expressed dismay at the variety and incomprehensibility of the tools that had been created. The instruments varied in their scoring techniques, substantive foci, and theoretical underpinnings.

Researchers had solved the problem of reliably ascertaining meaning by stipulating allowable meanings within each assessment instrument. Observers could be trained to see the particular meanings that researchers were interested in, so that there was interrater agreement within a research project even if different meanings were documented across research projects. Rosenshine and Furst (1973) demonstrate these differences in their contribution to the Second Handbook of Research on Teaching. They present a relatively common
classroom event in which a teacher poses a question and a student offers an unusual or unexpected response. The teacher then says, “That’s an interesting idea.” Unexpected student ideas are common occurrences in classrooms and typically present dilemmas for teachers. Teachers’ responses have been evaluated on many different observation forms, and Rosenshine and Furst show how this particular event would have been coded in a collection of different observation instruments. They take their sample of coding categories from instruments that were catalogued in Simon and Boyer’s (1967b, 1969, 1970) anthology of classroom observation instruments. Here are the relevant coding categories offered by different instruments, along with a reference to the instrument number within the Simon and Boyer anthology and the theoretical basis for the instrument. Rosenshine and Furst argue that all of these coding categories are intended to capture the same event:

1. Teacher entertains even “wild” or far-fetched suggestions (from Brown, #36, based on Dewey)
2. Routine agreement (from Aschner and Gallagher, #3, based on Guildford)
3. Teacher accepts or uses ideas of students (from Flanders, #5, based on a theory of interpersonal relationships)
4. Sanctions-search (from Joyce, #11, based on miscellaneous sources in instructional theory)
5. Exits-approved (from Medley, #13, originated by the author)
6. Evaluate without public criteria (from Miller and Hughes, #14, based on Hughes’ hypothesis about Group Processes) [Rosenshine & Furst, 1973, p. 146]

These different coding categories allow observers to document the fact that the teacher responded positively to an unexpected student idea, but each instrument gives a different meaning to the event. By training observers to see teaching in one particular way, each instrument achieves reliability.

By the mid-1980s, though, researchers began to believe that they had zeroed in on a set of teaching practices that improved student achievement. In the 1986 Handbook of Research on Teaching, Brophy and Good identified three broad categories of classroom practices that were by then known to be related to student achievement gains:

1. Quantity and pacing of instruction. Included in this aspect of teaching quality are things like the volume of content covered; the amount of time that students are academically engaged (which depends on the teacher’s ability to keep things running smoothly, keep transitions brief, and so forth); ensuring that students are continuing to be successful as they move along; and actively teaching rather than assigning student seat work.
2. Organizing the information that is presented. This includes structuring lessons with advanced organizers, providing some redundancy, and being clear in explanations.

3. Questioning the students. This factor includes a variety of features of teachers’ questioning practices, such as ensuring that questions vary in their difficulty but ensuring that most can be answered by most students; ensuring that the cognitive level of the questions is mixed; ensuring that questions are clear; and giving students time to formulate their answers and providing feedback.

The success of this body of research motivated many states and school districts to respond by embracing low-inference assessment systems that focused on these specific practices. They abandoned their rating scales and their interest in personality traits and adopted new low-inference techniques that focused on teaching practices rather than on the teachers’ personal qualities. This new research gave districts a way to settle their long-standing disputes about criteria for good teaching and at the same time to create more objective assessment instruments. The move to research-based teacher evaluation (RBTE) seemed to solve both assessment problems.

REPLACING PUBLIC VALUES WITH RESEARCH-BASED CRITERIA

The ideas from this body of research quickly migrated from the research community to school districts. There was a widespread sense that we had finally solved the problem of how to recognize good teaching, and the literature began to be populated by instrument developers describing, with pride, new administrative instruments that reflected research findings (Ellett, Capie, & Johnson, 1980; Peterson, Kromrey, Miccetti, & Smith, 1986).

However, almost as quickly as RBTE became fashionable, it became the object of criticism. Critics pointed out that most of the research findings came from lower grade levels and would not necessarily apply when teachers were teaching older students and more advanced content; they noted that intellectual life in the classroom had been overlooked (though it was also overlooked in earlier high-inference rating schemes) and that student achievement test scores were not an adequate criterion for evaluating teaching practices because they captured only a narrow range of the outcomes we sought from students. There were concerns that the findings encouraged administrators to prescribe and regiment teaching practices in a way that overlooked teachers’ need to adapt to variations in context and circumstances.

These problems with RBTE became especially evident when Greta Morine-Dershimer (1986), then AERA president, asked a group of scholars to write
critiques of a particular lesson, each using his own approach to evaluation. The lesson was a high school history lesson given by William Bennett, U.S. secretary of education at that time. The lesson certainly did not represent typical teaching as it occurs in the United States. One difference was that Bennett was there for just a single lesson on a single day; someone else taught the students the day before, and someone else would teach them the day after. The demonstration also took place in an honors class and was videotaped by all the major news networks. The camera lighting was so hot that the secretary was visibly sweating throughout the lesson. Students no doubt had been prepared by their regular teacher on how to behave when the secretary came.

In the lesson, Bennett taught primarily through a recitation-type examination of Federalist Paper No. 10, which offers an argument about the role of government based on a set of observations about human nature. Bennett presented the argument, as well as the view of human nature, as central to the foundation of the United States. Bennett had an intense personality. He paced the room and, pounding one hand against the other, pressed students to think harder about the Federalist paper they had read. Students appeared to be generally passive, uncertain how to respond to him.

The fact that this lesson was taught by a politician also hinders our ability to interpret it purely as an instructional event. I have presented a videotape of this lesson on more than one occasion in graduate courses on research methods, asking students to observe and document what they saw. I found that my American graduate students, who tend politically to be liberal Democrats, mostly noticed flaws and weaknesses in this lesson offered by a conservative Republican politician. Students from other countries, however, nearly always mentioned how much they learned from the lesson.

Since the lesson was such a politically charged and self-conscious event, we might expect that the best strategy for evaluating it would be to use an RBTE instrument, for such an instrument would ensure reliable and objective measurement. And in fact one of the responses to Bennett’s lesson (Peterson, Kromrey, Micerri, & Smith, 1986) applied the Florida Performance Measurement System illustrated in Exhibit 8.1. These authors offered an evaluation of the secretary’s lesson that gave norm-referenced scores for each of the following items, arranged here according to the scores Bennett received.

Bennett was in the top 25 percent on:

- Analysis/reasoning
- Concept treatment
- Corrective feedback
- Body behavior shows interest
- Express enthusiasm/challenge
He was in the middle 50 percent on:

- Handles materials orderly
- Orient/maintains focus
- Beginning/ending review
- Single factual question
- Specific academic praise
- Discuss cause/effect
- Value judgment
- Emphasize important points

And he was in the lowest 25 percent on:

- Begins instruction promptly
- State/apply academic rule
- Provide for practice
- Homework/seatwork assign
- Circulates and assists
- Stops misconduct
- Maintain momentum

These scores do seem to present a factually accurate representation of the secretary’s lesson. He was intense and focused on the central concepts in the essay, features that would lead to high scores on the enthusiasm and concept treatment. There was no time during the lesson when he stopped actively teaching in order to give students seat work exercises, so Bennett received low scores on variables associated with seat work.

Here, in a nutshell is the fundamental problem with the RBTE approach to teacher evaluation, for the items on which Bennett scored low were irrelevant to this particular lesson and there is no reason to believe that he should have given students repetitive practice in order to grasp the essence of the essay he was teaching. The authors concluded that the secretary of education had scored at the fiftieth percentile (presumably because he was high on the intellectual dimensions but low on the seat work exercise dimensions) and therefore would not be qualified to teach in the state of Florida. The rating was more impartial than those my doctoral students offered, but it still incorporated a bias toward a particular view of instruction.

Perhaps the most poignant response to Bennett’s lesson was from Barak Rosenshine (1986), whose comments centered on the discrepancy between measurable behaviors and intellectual importance. Rosenshine pointed out that research on teaching had focused on the teaching of skills (teaching students
to subtract double-digit numbers, for example) rather than on the teaching of content (the planetary orbits, the history of the United States, or mitosis and miosis). The kind of good teaching that had been revealed by low-inference observation instruments entailed presenting a skill, demonstrating it on the board, guiding students through some examples, giving them independent practice, and then circulating around to monitor their learning. But Bennett’s lesson was not a lesson about how to do something; it was a lesson about an idea. Rosenshine volunteered the remarkable admission that we researchers had not found a way to characterize the content that teachers teach. “We have not developed a technology for practice in this type of lesson,” Rosenshine admits, and then says, “Because there is relatively little research on the teaching of content, I find it inappropriate to critique this lesson” (p. 304). This is an astonishing concession and suggests that the preceding eighty or so years of efforts devoted to recognizing good teaching, in spite of the tremendous array of qualities it had examined, had nonetheless overlooked the most central aspect of teaching: helping students learn subject matter.

Research-based teacher evaluation, then, was unable to solve the century-long problem of establishing a set of agreed-on public values or of balancing objectivity with attention to the substantive meaning of classroom interactions. But criticisms of this approach laid the groundwork for attention to the intellectual work of teaching. From the ashes of low-inference, research-based teacher evaluation arose a new approach, often called a standards-based approach, which represents our current solution to the problem of accommodating public values and the problem of balancing objectivity with meaningfulness.

**THE STANDARDS SOLUTION**

The most prominent approach to performance assessment today is the standards-based approach, best exemplified by the National Board for Professional Teaching Standards and the Interstate New Teacher Assessment and Support Consortium, which focus on expert teachers and new teachers, respectively. Both groups organize their assessment around five central principles:

- Teachers are committed to students and learning.
- Teachers know the subjects they teach and how to teach those subjects to students.
- Teachers are responsible for managing and monitoring student learning.
- Teachers think systematically about their practice and learn from experience.
- Teachers are members of learning communities.
The tenor of these propositions is a strong and clear contrast to RBTE and a clear return to public values as the source of our criteria for good teaching. Whereas RBTE focused on a specific list of empirically documented behaviors, standards-based approaches emphasize professional knowledge, commitment, and judgment. Notice that each sentence begins with the word teachers, not the word teaching. Whereas RBTE focused on observed practice, standards-based approaches look at a broader range of evidence that includes artifacts from practice and teachers’ critiques of their own practice. Whereas RBTE offered a single set of behaviors for teachers in all grades and subjects, these standards-based systems are developing different assessments for different grades and subjects.

But the standards-based approach to measuring teaching quality did not abandon the concern with precise language and reliable measurement that characterized RBTE. The National Board’s procedures represent a substantial leap from the judgmental ratings that school administrators used in the early part of the twentieth century (see, for instance, examinations by Bond, 2000; Bond, Smith, Baker, & Hattie, 2000) in its effort to accommodate public values and balance attention to the meaning of events with attention to objective and reliable scores. Editor’s note: Chapter Ten provides more detail on the National Board assessment process and its rationale.

The National Board’s assessments also move substantially beyond a simple observation of teaching practice itself by examining teachers’ reasoning and rationales for their practices, attending to student learning as well as teaching practices, and giving explicit attention to the intellectual quality of teachers’ lessons. There is a cost for this success, however. The board’s assessment system is a lengthy and costly affair. Costs are prohibitive for many teachers, and critics have raised questions about the cost-benefit ratio. It is not intended to be, nor could it be, a system that school districts could feasibly incorporate into their annual performance evaluations of teachers. But it does appear to provide a reliable method of recognizing good teaching when we see it.

Districts prefer more cost-efficient approaches to standards-based assessments, and in the past decade, standards-based assessments of teaching have almost completely replaced RBTE throughout the education system. Probably the most widespread approach to standards-based teacher evaluation derives from Danielson (1996), who offers a framework that districts can adopt wholesale or use as a starting point for devising their own system. Danielson identifies four major domains: planning and preparation, classroom environment, instruction, and professional responsibilities, with each domain containing itemized lists of specific standards within it. Moreover, Danielson claims her system is based in research. So even if specific evaluation forms vary from district to district, the terms are not precise, and the reliability of ratings is unknown, we could consider her standards-based evaluations
to be a research-based approach. Editor’s note: See Chapter Six for how the Danielson Framework is incorporated into Induction programs.

However, the Danielson system includes a number of items that look more like statements of public values than like research-based findings: items such as “integrity and ethical conduct,” “service to the school,” and “relations with colleagues.” In fact, this system tries to incorporate both fashionable ideas and enduring ideas into a coherent and practicable system. In that sense, it serves a purpose very much like Boyce’s 1915 guidelines did: a consensus overview of things we care about. Moreover, the phrasing of these standards, and many other locally developed standards, is such that they require relatively higher inferences on the part of the evaluator, who needs to evaluate things like “quality of questions,” “suitability” of learning goals for diverse learners, or “accuracy” of reflections on teaching. Still, Danielson’s system offers a substantial improvement over earlier judgmental rating scales, and there is some evidence that district-level teacher assessments based on the Danielson framework are associated with student achievement gains (Gallagher, 2004; Kimball, White, Milanowski, & Borman, 2004; Milanowski, 2004).

The standards-based approach to teacher evaluation is an improvement over the early rating scale and an improvement over RBTE. Its focus on standards reintroduces public values and provides a language for helping districts sort out the values they care most about. Its reliance on performance rubrics increases the objectivity and reliability of scores. But the instruments currently in use by local districts are still remarkably various in the aspects of teaching they evaluate, and many still lack sufficient reliability as well. Next I examine the current state of teacher evaluations.

Aspects of Teaching Currently Evaluated

Although we tend to speak of standards-based evaluations as if they represented a single approach, there is tremendous variety among the standards-based instruments used by school districts and by researchers. Recently I analyzed the contents of four illustrative teacher assessment instruments (Kennedy, 2007). Two were instruments used by school districts to evaluate their teachers, and the other two were used by researchers to study teaching. I sorted the items in these instruments into the following categories:

- Professionalism: teachers’ planning, record keeping, out-of-classroom contributions to the school, interactions with parents and others
- Organization and ambience of classroom life
- Inclusiveness in discourse and classroom activities
- Clarity of communications within the classroom
- Thoughtfulness of classroom discussions
- Quality of the students; learning activities
These categories can be viewed as arrayed in a continuum from those most distant to those most immediately relevant to student learning. The first category, professionalism, includes teaching practices that are most distant from student learning: things teachers do when they are outside the classroom. The next two categories have to do with the general organization of the classroom and the treatment of students, things that are important but not directly related to student learning. The fourth category, clarity, is the first category in the list that we might expect to have a direct bearing on learning. The last two categories, those having to do with the kind of intellectual work students are doing, would presumably bear most directly on their learning. These are aspects of teaching that were largely absent from teacher evaluations in the early part of the twentieth century and in the RBTE instruments but that are now beginning to appear in standards-based instruments.

Figure 8.1 displays the percentage of evaluation items in each assessment system that addressed each of these six broad public values. The first two instruments came from school districts and the second two from researchers. Notice that just as we did a century ago, we still face substantial differences in how much attention we give to these different criteria. For example, the first instrument gives no attention to the nature of the learning activities students participate in, whereas the third and fourth give substantial attention to that. On the other side, the first and third give substantial attention to professionalism, while the fourth does not address this at all. These differences do not necessarily reflect a disparity between district and researcher either.

Figure 8.1 Content Analysis of Four Teacher Assessment Instruments
Notice that inclusiveness, the third category, receives no attention in the first system but quite a bit of attention in the second. Thus, a given teacher would be evaluated against quite different criteria depending on which of these districts employed him or her.

The second important message from this figure is that the school district systems give the bulk of their attention to aspects of teaching that are relatively more distant from student learning. These two assessment systems were selected because they represent common approaches to school district teacher evaluations, but in fact, both lack attention to the very things that constitute the intellectual core of teaching: classroom discussions and learning activities. Even after a century of efforts to improve teacher evaluations, most district evaluations still overlook the intellectual work that occurs in classrooms. No doubt one reason for this omission is that these are difficult things to evaluate, and we are only now beginning to devise solutions. This observation brings us to the second problem, which still remains to be solved.

**Objectivity in Scoring**

The two assessment samples shown in Exhibits 8.1 and 8.2 illustrate the difference between the high-inference, judgmental ratings used at the beginning of the previous century and the low-inference documentation of practice used for research-based teacher evaluations. Standards-based systems use a third approach, performance rubrics, which lie in between these other two approaches. Typically a performance rubric is based on a scale, like the rating scale shown in Exhibit 8.3, but instead of asking for a judgment about how far to the left or right a teacher is, the scale defines what an observer might see if the teacher were to be assigned a score of a “1” or a “2.” Still, the phrasing of these items can vary substantially in how much judgment the observer is expected to draw on in establishing a rating. Exhibit 8.3 illustrates a handful of performance rubrics, all focused on classroom teaching practices. They are listed in order of their dependence on the observer’s subjective interpretation of events. Those at the top are higher-inference items, those at the bottom lower. The first one, for instance, asks whether “instruction respects student knowledge.” Deciding that a teacher’s practice respects student knowledge is a highly interpretive act. In addition, the item gives only a general idea of how the points on the scale differ from one another, so judgment is required here as well. The second item is only slightly better. It offers descriptions of each point on the scale but still depends on the observer to evaluate the appropriateness of the teacher’s strategies.

As we move down this list, item scores are defined more and more precisely, thus depending less and less on the judgments and interpretations of the observer. The last item asks about the proportion of students who participated in a discussion and gives relatively precise descriptors of the participation rates.
## Exhibit 8.3 Illustrative Performance Rubrics

<table>
<thead>
<tr>
<th>Lowest precision</th>
<th>Instruction Respects Students’ Prior Knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. never 2. sometimes 3. often 4. typically</td>
</tr>
<tr>
<td><strong>Instructional Strategies</strong></td>
<td></td>
</tr>
<tr>
<td>Lowest precision</td>
<td>1. Are ineffective or inappropriate 2. A limited range is effective and appropriate to the content 3. Typically are effective and appropriate 4. Broad range of effective and appropriate strategies</td>
</tr>
<tr>
<td><strong>Teacher Questions</strong></td>
<td></td>
</tr>
<tr>
<td>Low precision</td>
<td>1. Generally poor quality 2. Combination of low and high quality but rapidly given 3. Most are adequate to high quality, and time is given for response 4. Uniformly high quality, with adequate time for response</td>
</tr>
<tr>
<td><strong>Learning Activities</strong></td>
<td></td>
</tr>
<tr>
<td>Medium precision</td>
<td>1. Students repeat or recall declarative knowledge 2. Students apply a known procedure; may be asked to show their work 3. Something in between 2 and 4 4. Students use complex reasoning or apply nonalgorithmic thinking</td>
</tr>
<tr>
<td><strong>Response to Student Errors</strong></td>
<td></td>
</tr>
<tr>
<td>Higher precision</td>
<td>1. Does not correct student errors 2. Corrects errors but without explanation 3. Corrects with an explanation that clarifies the process or concept 4. Corrects with an explanation that clarifies and also facilitates student self-correction</td>
</tr>
<tr>
<td><strong>Classroom Participation</strong></td>
<td></td>
</tr>
<tr>
<td>Highest precision</td>
<td>1. Less than 25% of the students participated in the discussion 2. Twenty-five to 50 percent of the students participated minimally in the discussion (they contributed only once) 3. Twenty-five to 50 percent of the students participated consistently in the discussion OR over 50 percent of the students participated minimally 4. Over 50 percent of students participated in classroom discussion</td>
</tr>
</tbody>
</table>
required for each score. The one just above it asks whether or how teachers respond to student errors. Here, four scale scores are possible, and each is associated with a uniquely defined response.

These performance rubrics are a substantial improvement over the undefined rating scales of the previous century, and a substantial improvement over RBTE scales as well. They remove some of the judgment that early rating scales encouraged, but at the same time they acknowledge that teaching practices are meaningful events. But the variations shown in Exhibit 8.3 suggest that we still have far to go. One reason teacher evaluation scales have not improved more quickly is that the articulation of these distinctions requires a great deal of thought and observation. Distinguishing between better and worse approaches turns out not to be self-evident at all, and in fact requires close attention to, and analysis of, details.

**DISCUSSION**

We have accumulated a century of experience trying to systematically recognize good teaching when we see it. Although most students and parents believe they can spot good teaching in an instant, education administrators and researchers have struggled to develop reliable procedures for distinguishing among teachers. One reason for the difficulty of this task is that we have faced serious technical problems in trying to translate our intuitive perceptions into precise terminology and reliable measurement scales. In fact, it has taken almost the entire century to distinguish the teacher from the teaching.

Early instruments routinely evaluated the entire person and focused on such personal qualities as dress, comportment, loyalty, and efficiency. Even efforts to focus on classroom practices still tended to look at subjective qualities such as friendliness or orderliness. Our century has been filled with public hand-wringing by researchers and school administrators who struggle with the difficulty of obtaining fair, accurate, justified evaluations from observers with different personal or theoretical views about teachers and teaching.

Associated with this problem of translating intuitive perceptions into standardized protocols is the fact that until recently, assessments have not attended to the intellectual substance of teaching: to the content actually presented, how that content is represented, and whether or how students interact with it. This may seem like a surprising and glaring omission, especially since it has been pointed out more than once (Buchmann, 1982; Rosenshine, 1986; Shulman, 1986). But it is not surprising at all when we consider the difficulty of agreeing on the meaning to the events we see. And any assessment of the intellectual and substantive merits of teaching is entirely about its meaning. The central problem here has to do with the difficulty of developing reliable procedures
that can capture the substantive meanings of events rather than just the events themselves, and yet do so in a way that is not idiosyncratic.

The most significant problem remaining in most state and district assessments of teaching, including those based on Danielson’s *Enhancing Professional Practice: Framework for Teaching*, is that they still do not give sufficient attention to the substantive and intellectual merits of classroom lessons themselves. Evaluators may be asked to judge the clarity of lesson plans, the quality of classroom interactions, the physical arrangement of the classroom, or interactions with parents, but they rarely are asked to evaluate the accuracy, importance, coherence, or relevance of the content that is actually being taught or to rate the level of intellectual work students are asked to do. This absence of attention to content is especially puzzling in Danielson’s Framework because Danielson cites many chapters from Wittrock’s 1986 *Handbook of Research on Teaching* as the source for the elements in her framework. Yet it was a chapter in that very handbook (Shulman, 1986) that first brought to our attention the need for more attention to content in research on teaching.

Documenting the intellectual meaning of teaching events remains the elusive final frontier in performance assessment. Researchers are now moving in this direction and are exploring a variety of approaches to defining the substantive meaning and intellectual value of classroom lessons. They are finding ways to characterize the logic of whole lessons (Stigler, Gallimore, & Hiebert, 2000), the intellectual requirements entailed in student assignments (Newmann & Associates, 1996), and the content and thoughtfulness of classroom discourse (Resnick, Matsumura, & Junker, 2006). These new instruments are often limited to particular subjects and grade levels, and they require more observer training than school districts may be willing to embrace, but they offer districts a glimpse of their own likely future.

Mary Kennedy appreciates financial support from the National Science Foundation (NSF). This book was motivated and influenced by NSF grant award no. 0554477. Any opinions, findings, and conclusions or recommendations expressed here are those of the various authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES


