Teaching and Learning: The Journal of Natural Inquiry & Reflective Practice

Volume 21 | Issue 2 Article 6

1-2007

Making the "A" List: Negotiating Assessments, Standards, and Teacher Development

Todd W. Kent

Teaching & Learning, Spring/Summer 2007, Volume 21, Numbers 2 & 3, pp. 194-212

Making the "A" List: Negotiating Assessments, Standards, and Teacher Development

Todd W. Kent
Princeton University

Creating an assessment that captures all the complexities of teaching has been a difficult and elusive feat for the field of education. As in the case of distinguishing good art from bad, we all recognize the differences between good and bad teaching when we see or experience it. However, as with art, there is a degree of subjectivity in valuing the quality of teaching that reflects the experience and expertise of the evaluator. Most parents are certainly willing to tell you who the good teachers are in a school and which teachers are best avoided, yet what appears to be "good" teaching to one person may be valued quite differently by another. One parent may describe a horrible yearlong ordeal for his or her child with a teacher who turns out to be perfectly acceptable for your own. What determines "good" and "bad" varies greatly between individuals because of differences in values and differences between the needs and characteristics of their children.

Each semester I give a "Three Teachers" activity to the students in the Seminar on Learning and Teaching, the introductory course for the Program in Teacher Preparation at Princeton University that I co-teach with Helen Martinson. The activity, borrowed from a workshop hosted by Alverno College on the Interstate New Teacher Assessment and Support Consortium (INTASC) standards of teaching, asks students to rank three teachers based on written descriptions taken from portfolio evaluations of their teaching. Each semester, students have little problem discriminating between the three teachers and reaching a consensus on the ranking of the three from best to worst. When we ask them to provide the criteria they used to distinguish levels of quality, we find that these characteristics nearly always correspond to the descriptions of the ten INTASC standards. The conclusion the students draw from this exercise is that they are able to develop criteria that describe good teaching, and they essentially recreate the INTASC standards by working through this task. What becomes problematic, however, is using those criteria to assess. Each semester during the same exercise, there are always

one or two individuals out of a class of eighteen who disagree with the rankings of the majority of their peers. This inconsistency does not arise because of disagreement over the criteria; disagreements usually occur over how much value should be given to specific criteria.

It is impossible to assess concurrently all the variables associated with the complex act of teaching. Students, parents, administrators, teacher educators, and government bureaucracies tend to focus on the variables that matter most to them in making their respective judgments about teaching and to use the variables they can measure most efficiently. Differing values result in subjectivity and inconsistency, and the field of education will always struggle to create consistent measures that can discriminate between levels of quality for teaching. The problem arises in the inevitable gap between what we *can* measure reliably and what we value and would actually *like* to measure. We can all but eliminate subjectivity in our assessments by creating carefully structured instruments that prove reliable and consistent over time and with large numbers of individuals. The price we often pay for such consistency, however, is the narrowing of what it is we are actually measuring. The result, then, is that we tend to measure selective elements of teaching or learning outcomes. In acknowledging this level of complexity, many teacher preparation programs turn to portfolios to collect various pieces of evidence that, when taken together, provide a somewhat comprehensive representation of a teacher's work and the impact of that work on students. The challenge, of course, is evaluating and interpreting a complex collection of artifacts.

This paper describes how the Program in Teacher Preparation at Princeton University approaches the problem of collecting and evaluating evidence that represents multiple facets of teaching. This discussion does not present the program's evaluation process as a model that others should adopt. Rather, the intent is to begin a frank and open discussion about the challenges of teacher evaluation and to describe how the program's context helped to shape its evaluation and assessment procedures and instruments. The program's context is partially defined, of course, by the current political climate that is preoccupied with the use of "data" and for evaluating teaching and program performance. The paper will describe how the program negotiates the tension between

the procedures for collecting data and the process of valuing and interpreting those data in making judgments about the quality of teaching.

Summative Assessment: Using Quality Measurements to Rank and Sort Students

One of the purposes of evaluation is to rank and sort students according to levels of performance. Programs of teacher preparation have an ethical responsibility to safeguard the profession from mediocre or poor performance, and this task requires summative evaluations of students for the purpose of determining whether or not they perform at levels that warrant their entry into the teaching force. State governments do not leave such determinations solely to individual programs, and most states require some sort of standardized examination that must be passed as a prerequisite for licensure. In New Jersey, teacher candidates must take the PRAXIS II test, administered by the Educational Testing Service (ETS), in their certification content areas. The PRAXIS II assesses only subject area knowledge, and the state sets the passing level for individual tests. Students must earn scores above these levels to be eligible for teacher licensure. In a sense, the tests are pass/fail; students are either eligible for certification or not, depending on the score they earn. Recently, ETS has issued certificates acknowledging candidates who have scored within the top 15% of test takers. Evaluation of a candidate's pedagogical and instructional skill is appropriately left to individual programs during preservice experiences and to schools during a candidate's first year in the profession. The PRAXIS II serves as a safeguard, ensuring that teacher candidates are adequately prepared in their subject areas, and the federal government recently decided to use the performance on the state standardized testing requirements as the primary measure of the quality of a candidate's preparation.

Title II, Section 207 of the 1998 federal Higher Education Act (HEA) requires all teacher preparation programs, starting in 2000, to report their pass rates on tests that states require for certification. The state collects these scores and then divides the programs into quartiles according to their pass rates. The Title II website states that the "ranking by pass rates is an incomplete measure of the relative quality of a state's teacher preparation programs," and for this reason the guide emphasizes

that institutions of higher education (IHE) and states may report supplemental information that they believe will help the public "better understand the relative quality of each IHE's teacher preparation program" (Title II, 2005, ¶ 3). In the case of New Jersey, the "supplemental information" consists of an overview of teacher preparation in the state including the New Jersey Department of Education (DOE) mission statement, a brief description of a comprehensive review of the administrative code, a brief statement explaining that the levels of qualifying scores have been reviewed and raised, a three-paragraph overview of the 21 IHEs that provide teacher preparation, and other information relating to national accreditation policy and DOE initiatives. The overview of teacher preparation explains that pedagogy is assessed during the teacher's first year of teaching because New Jersey candidates initially receive a provisional certificate that becomes a permanent certificate after the first year of teaching and that more than 99% of the teacher candidates prepared by New Jersey programs are successful in the classroom during this first year. The overview also provides some descriptive information, such as the number of public and private higher education institutions that prepare teachers in the state. This supplemental information supposedly provides the appropriate context for helping the public to interpret the rankings of programs by the pass rates of their students.

In 2003, the summary PRAXIS II pass rates in New Jersey ranged from 100% to 93%. The first quartile had a mean of 100%, the second quartile a mean of 98.3%, the third quartile 96.7%, and the fourth 94% (Title II, 2005). Thus, a scant 6 percentage points separates the "top" performing schools from the "lowest" performing schools. The Title II website states that the purpose of this system of ranking is to provide the public with "a clear and comprehensible public reporting system on state licensure and the success of institutions in preparing teachers" (Title II, 2005, ¶ 2). Thus, the purpose of these rankings, along with the supplemental information, is to enable the public to make judgments regarding the "success" of particular institutions. The federal government evidently prioritizes the values of "clear" and "concise" over such values as "complex" or "comprehensive." Certainly, the evaluative information presented by Title II is exceedingly narrow and captures only one small element of teaching quality—content knowledge as measured by the PRAXIS II. The irony of this

whole endeavor is that, for the most part, the content knowledge of teacher candidates is developed within colleges of arts and sciences, not in teacher preparation programs. Thus, teacher preparation programs are being ranked according to a quality of teacher preparation that is largely beyond their control. A second irony involves the narrow range of percentage points that separates the top performing schools from the "lowest" performing schools. The teacher preparation program at Princeton is very small; we typically certify between 20 and 25 students each year. Princeton currently enjoys a 100% pass rate and is therefore ranked in the first Title II quartile. If only one of our students had failed the PRAXIS in 2003, Princeton would have fallen from the first quartile to being dead last in the state. The running joke among our students is, naturally, "Don't be the one!" Each state has its own testing requirements, but the situation in New Jersey certainly demonstrates the inherent dangers of placing inappropriate emphasis on simplistic one-size-fits-all measures of complex processes.

The problem with Title II is not the PRAXIS II. The test is developed and administered by ETS, the quintessential force in standardized testing, and the company rightly vouches for the validity and reliability of its tests. The test is objective and consistent. The problem with Title II is that public rankings of "success" in teacher preparation are being compiled based on a single limited measure, a seemingly inevitable policy reality when comparative measures are desired across a broad population. To rank something as complex as the quality of teacher preparation by such a limited measure raises legitimate concerns, but Title II is nevertheless taken very seriously by teacher preparation programs because rankings are indeed formulated and made public. Title II serves as a prime example of trying to use what we *can* easily measure—the PRAXIS II is the *only* assessment taken by *all* teaching candidates in the state—to evaluate something complex that we *want* to measure: success in preparing competent teachers. A complex activity necessarily requires complexity in evaluation. No single piece of evidence or measurement can or should be used to represent complex performance.

Grant Wiggins, a leading voice in educative assessment, emphasizes the importance of gathering a range of different types of evidence in order to demonstrate understanding. In Wiggins' (1998) "backwards design" framework, understanding is a complex entity,

and he describes six facets of understanding to help guide educators in shaping instruction and assessment. His framework encourages educators to focus units of instruction on a few "enduring understandings" and to develop educative experiences that allow students to deepen their understanding while producing a range of evidence to determine whether the intended understandings had been developed. Such evidence can range from informal evaluation during instruction to simple quizzes and tests targeting knowledge and skills required for understanding to complex performances that indicate depth of understanding and the ability to apply that understanding to concrete tasks. The Wiggins system also relies on rubrics to describe and distinguish between levels of performance (Wiggins, 1998). These same principles certainly apply to the evaluation of teaching, but even complex assessments with carefully designed rubrics can be problematic.

A complex, rubric-based, evaluation system is used by Princeton's Teacher Preparation Program staff to rank individuals who have been nominated for the Distinguished Secondary Teacher Award, a prestigious honor bestowed upon four New Jersey teachers each year at the University's commencement ceremony. Discriminating among these individuals is exceedingly difficult because each represents "the best" teaching within an individual school. These are all accomplished individuals achieving remarkable feats in the classroom, in their schools, and in the profession. The nomination process requires each school to compile a dossier that presents the case for the nominee. The program staff reviews the dossiers of approximately 80 nominees that are submitted each year and selects the ten or twelve strongest for the second phase of evaluation which entails classroom observations and interviews. A committee of university staff and faculty and representatives from local schools then reviews the dossiers and school visit reports to select the four winners.

The program staff developed a comprehensive rubric and corresponding scoring form to rank the nominees for the first phase of the evaluation. The scoring form allows points to be assigned for each category of the rubric, covering such areas as the nominee's effectiveness as a teacher of the subject matter and his or her intellectual leadership among colleagues. There are six categories for evaluation. Complicating the task of identifying fine distinctions among a group

of talented and accomplished individuals is the fact that the nominees are drawn from an incredibly varied array of instructional positions and school environments. This past year's 76 nominees included, for example, an AP history teacher from a prestigious independent school, a computer teacher from a vocational school, a ballet teacher from an urban school, a middle school science teacher, and several Latin teachers. Creating a rubric that can compare the quality of teaching in such disparate circumstances is difficult at best. After applying the rubric for the first time in 2004, the staff decided to add two additional categories. The first new category covers "Special Considerations" that allow credit to be given to teachers who are working in unique circumstances or who have made distinctive contributions to schools or students that extend beyond the realm of "normal" teaching. An example might be a teacher who created or revitalized an entire area of study within an impoverished school district. The second new category allows the reviewer to assign a modest number of "Personal Quality Points" if the reviewer decides that elements of a dossier are particularly compelling. Essentially, these last two categories are provided to address complexity and circumstances that cannot be predicted but should, nevertheless, be considered.

We tabulate our individual scores to select the twenty or so teachers who receive the highest ratings from program staff, and then we meet for the better part of a day to discuss each teacher in detail. Although the staff developed the rubric and procedures for evaluating the dossiers, there can be wide variation in the ratings of some nominees. The rating correlations between individual assessors and the entire group of raters range from .55 to .80, and there can be high agreement on the ratings of some nominees as well as wide disagreement on others. The average range of scores between the highest and lowest raters was 16 points on a 75-point scale. Some nominees were scored with as little as a 7-point difference between the highest and lowest raters while one nominee had a substantial 27-point difference between high and low scores. Interestingly, none of the teachers with the narrowest (7 points each) or widest (27 and 26 points) spread of scores made the final selection of ten teachers. The average scoring range for the ten selected teachers was 13, indicating that there was a slightly higher level of agreement between raters on the "best" nominees than with the group overall. Although our staff is comfortable using the rubric

scores to help identify the top third of the dossiers, none of us would be comfortable relying on the scores alone to determine the final ten or so teachers selected for the second phase of evaluation. We have found that the discussion among our five staff members is essential because rubrics and scores are necessarily limited in what they capture and are therefore inadequate for making fine distinctions between individuals who have all attained a high level of quality in the evaluated area.

During the staff discussion, scores can be challenged, defended, or adjusted as information is reinterpreted or new perspectives are considered. For example, two individuals may have similar levels of accomplishments within the classroom and in school life, but one teacher may have stayed in the same school for thirty years while the second, younger teacher, has changed jobs every three to five years. Whether, or how, differences in employment patterns or length of career should be considered as evaluation criteria is beyond the capability of the rubric and can only be addressed through thoughtful discussion and by considering specific circumstances and professional context for each individual. Partially by design and partially through trial and error, we have developed an evaluation process that has intentionally integrated some subjectivity into the scoring process that results in less consistency between raters but allows for consideration of the wide variation we find among the nominees. This subjectivity is resolved with collective professional judgment through thoughtful discussion. We are evaluating the "best" teachers from schools across the state, and the staff takes this responsibility seriously. Collectively, we are much more comfortable with this hybrid form of evaluation than if we had worked to refine the rubric and scoring sheet to improve the consistency of scores between our staff. To do so, we feel, would mean narrowing the criteria to the extent that we would not be able to consider the "whole" teacher. The addition of two "catch all" categories to the process created flexibility to more effectively discriminate between widely ranging teaching situations. The staff discussion, structured by the rubric and scoring procedures, allows us to apply professional judgment to the interpretation of individual scores. The scores provide an essential and useful common measure, but relying upon the scores alone without some process to interpret the meaning of those numbers would not yield our current level of satisfaction that we truly select the "best" dossiers.

We essentially use this same process of blending "objective" measures with collective interpretation when evaluating the clinical performance of our teaching candidates. There are significant differences between the two processes, but perhaps the most important is that we use the New Jersey Professional Standards for Teachers as the primary rubric for evaluating teacher candidates. A second important difference is our ability to collect a wider range of evidence over a longer period of time for our students than we can with the teaching award nominees. A third significant difference is that the discussion involving our students is ongoing during practice teaching. During the course of a semester, our staff meets every other week to discuss in detail the progress and challenges of each student who is practice teaching. At the end of the semester, the staff then meets to discuss the practice teaching grade of each individual.

Grading decisions for students completing their practice teaching reflect a range of evidence. This collection of evidence includes interim and final teaching evaluations rated by the three individuals directly involved in the practice teaching process: the university supervisor, the cooperating instructor, and the student teacher. These evaluations are supplemented with a submitted Work Sample, observation reports, journal entries, and anecdotal information. Student teachers, cooperating instructors, and university supervisors all use the same standards-based form to evaluate student teaching performance, and these evaluations are considered the most compelling evidence of a candidate's teaching skill. Student teachers can be rated for a given criterion as "Targeted for Improvement," "Proficient," or "Exemplary." An analysis of these scores indicates that supervisors and cooperating instructors were very close in their scoring, giving exemplary scores an average of 38% and 36% of the time, respectively. Student teachers, however, were much harder on themselves, giving exemplary scores only 19% of the time. The potential for discrepancies among the three raters highlights the importance of establishing a process of evaluation that can account for a wide range of performance indicators and variations between those measures. The grade discussion of each student begins with the supervisor describing the overall performance of the student, performances on the individual assessments, and a proposal for a course grade or range of possible grades. Each member of the staff has the opportunity to discuss the grade proposal,

to ask questions about performance, or, if necessary, to advocate for the student. Because we regularly discuss all candidates during their student teaching and because so many of our staff have interactions with most of our students (as course instructors, as advisors, or as supervisors), the discussion of any particular student will involve most of the staff. This process of interpreting the entire range of evidence also allows for the consideration of unique circumstances, such as personality clashes between a student and the cooperating instructor or university supervisor. As in the case of the Distinguished Secondary Teacher Award evaluations, data are considered but are interpreted with the application of professional judgment through thoughtful discussion.

Formative Assessment: Helping Students to Meet Standards

In January of 2004, New Jersey adopted new licensing regulations for teachers. The new regulations contained the New Jersey Professional Standards for Teachers, a set of ten standards based on the INTASC general standards. Like the INTASC standards, the ten New Jersey standards are each broken down into the areas of knowledge, values, and commitments (called dispositions by INTASC) and activities (called performances by INTASC). The adoption of standards was a very significant decision for New Jersey because teacher quality was now defined in terms of what teachers know, believe, and are able to do. In the previous iteration of the New Jersey licensing code, teacher quality was defined in terms of required academic experiences, based upon the Boyer Topics, for teaching candidates. Although the new licensing code also contains proscriptive requirements governing the content of teacher education curricula, the shift of emphasis from coursework to performance-based standards was a welcome development, especially for a small program like the one at Princeton.

In Fall 2002 and in anticipation of the new licensing code, Princeton's Program in Teacher Preparation staff began the process of completely aligning the program elements with the INTASC standards. We began by reviewing every component of the program to identify whether or not the standards were being addressed and to determine whether or not artifacts were being generated that could be used as evidence that students were meeting the standards. This process led to substantial revision of program components and to the alignment of

student teaching evaluation instruments to the standards. When New Jersey passed its own standards, these changes were adjusted to reflect the relatively minor differences between the two sets of standards. The transition to standards facilitated revisions that increased the coherency and consistency of the various program elements while preserving the unique character of the program.

Our small program has succeeded over the past 35 years because we give students a great deal of individual attention. We typically certify between 20 and 25 students each year, spanning all content areas and all levels of teaching. Students fulfill state requirements by taking courses taught in our program and in other academic departments. Our students can be majors in any department at the university, and they fulfill our program requirements along with their departmental and general education requirements. We do not have the luxury of being able to require our students to take a large number of education courses, as is the case with programs embedded in larger schools of education. We must compensate for fewer course offerings and fewer contact hours with our students by individualizing the program to students. When New Jersey moved toward a standards orientation in their licensing code, our program seized the opportunity to reinvent itself with the goal of developing teacher candidates who perform skillfully in the classroom as defined by those standards. Our program philosophy is to move each student to that level of performance by differentiating program instruction and providing individual attention for each student. As we revised our program, the staff decided that a teaching portfolio was the best vehicle to help our students meet the new standards.

When a student is admitted to the program, he or she meets with our director, John Webb, for the Introductory Practicum. This program requirement is a non-credit experience that serves as an introduction to the program and to the standards with which we are aligned. The student is given a CD-ROM that contains all the New Jersey Core Curriculum Content Standards, the Professional Standards for Teachers, and Subject Specialty Standards for the student's area of certification. The student is also introduced to an observation form (used by student teachers, cooperating instructors, and university supervisors during practice teaching) based upon the Professional Standards. The student is then given a second CD-ROM containing

the video of a lesson taught by a past student teacher, and the student views the video through the lens of the standards. After completing this "training" exercise, the student again meets with our director to discuss and interpret the video from the perspective of the standards. After this in-depth, one-on-one discussion, the student is placed with a teacher in a neighboring school district for a day of observation. Again, the standards provide the lens for processing what is observed that day, and the student returns for a final discussion to debrief this experience. This introductory exercise is fundamentally important to our program in that it introduces all the relevant standards and because it allows each student to examine teaching in light of the standards. The in-depth, tutorial nature of this experience allows our director to ensure that the students begin the program with an orientation that serves as an advanced organizer for the rest of the program. The experience also allows the director to identify and develop the relevant dispositions of students as well as address any misconceptions or detrimental preconceptions that might surface and inhibit the learning of the student.

Upon the conclusion of the Introductory Practicum, the director assigns an advisor to the student. The primary role of the advisor is to oversee and assist the student in the development of the portfolio. Students are required to submit two to three artifacts, or pieces of evidence, of their competency for each of the three areas of knowledge, values, and activities under each of the ten Professional Standards. Our program recognizes that students may take many different pathways to the same end: professional skill as defined by the standards. The portfolio provides the means to truly individualize instruction. Students self-assess their portfolio in anticipation of each of the five advisor meetings scheduled for each academic year, and this assessment serves as the starting place for the advisor meeting. There are required elements from the program that must be represented in the portfolio, but students have considerable latitude in deciding which specific pieces of evidence they would like to submit. If there is an area of weakness identified in the portfolio, advisors might recommend readings or other educative experience to address the area. In this way, the portfolio allows the program to both assess and develop on an individual basis the student knowledge, dispositions, and performance defined by the standards.

Most of the assessment given to students during their practice teaching is formative. It is our goal, as a program, to raise every student to the "A" level of teaching skill, as defined by the New Jersey Professional Standards for Teachers. Although the cooperating instructor provides regular feedback in the form of observation reports and the interim and final evaluation forms described above, we consciously avoid bringing the cooperating instructor directly into the grading process to avoid interfering with the mentoring relationship that develops between the host teacher and the student teacher. The evaluations by the cooperating teacher are included among the many pieces of evidence considered for each student during our grading meeting at the end of the semester. We make it clear to the students that the university supervisor, representing the program, has the sole responsibility for the grade.

The supervisor does not formally "grade" the student teacher during the weeks of practice teaching. Our program chooses to base the practice teaching course grade on a body of evidence considered in entirety at the conclusion of the practice teaching experience. The supervisor writes up six formal observations of each student during practice teaching and meets with the student for about an hour to discuss each observation. These observations serve primarily as opportunities to provide feedback to the student and to prompt student reflection and self-adjustment, rather than to generate a grade. The interim evaluation is also considered formative, and the supervisor truly serves as a coach rather than evaluator during the practice teaching experience. The formative nature of the student teaching experience is designed to help each student evolve into a competent and skilled teacher, as defined by the standards.

Self-assessment by the student is also an important part of this process. The student fills out the same interim and final evaluation forms that the cooperating instructor and university supervisor use. Each student is videotaped twice during practice teaching. The first taping is done early in the placement and provides an opportunity for the student to view him- or herself in the process of teaching. The supervisor provides an observation report on this taped lesson, enabling the student to have an experienced educator's perspective of the lesson. The students are then videotaped a second time during the second half of the placement. The student, and not the supervisor, responds to this

second taped lesson using the same standards-based observation form that the supervisor used for the first taped lesson. The student's response to the second lesson is turned in as part of the student's Work Sample and is included in the range of evidence, to represent both teaching skill and the student's ability to reflect upon his or her own performance, when considering grades at the end of the semester. Students also keep a daily journal during practice teaching. The journal provides the opportunity for reflection by the student and also provides the supervisor with some description of events between scheduled observations. The supervisor provides responses to journal entries, and though not graded, the entries provide some indication of a student's ability to informally reflect and self-assess.

The grades assigned for student teaching serve as a means for conveying quality to potential employers. Student grades reflect the program's judgment of teaching skill, defined by the standards, but the program must also adhere to university guidelines on grading. On April 26, 2004, the Princeton University faculty voted by a two to one margin to approve a new policy setting universitywide guidelines limiting the number of A's (defined as A+, A, or A-) awarded by departments or programs to fewer than 35% of the grades given to students in undergraduate courses. The impetus for this new policy was to combat grade inflation, a growing problem that Princeton shares with most other institutions of higher education. Why did Princeton take such a bold step toward curbing grade inflation? In a letter to parents, the Dean of the College, Nancy Weiss Malkiel (2004), who proposed the grading guidelines to the faculty, described the reasoning this way: "We think it is important to address grade inflation because of the way it affects teaching and learning at Princeton" (p. 9). The Dean went on to describe the following four assumptions about grading and assessment:

1. Grading, properly done, is an educational tool that assists students in evaluating what they have learned, how well they have learned it, and where they need to invest additional effort.
2. Grading done without careful calibration and discrimination is, if nothing else, uninformative and therefore not

useful; at worst, it actively discourages students from rising to the challenge to do their best work.

3. Students are entitled to a fair and reasonable assessment of the work they have done; there should be some correlation between performance and reward.
4. It does students no favor to grade them in a way that fails adequately to differentiate routinely good from really outstanding performance. We need to do a better job of distinguishing the excellent from the competent and of holding students accountable for negligent, weak, and unacceptable performance (Malkiel, 2004, p. 9).

Princeton's new grading policy, and these four underlying assumptions, have implications for teaching and learning in any academic area, but they pose specific challenges and considerations for a program that prepares teachers. The first assumption describes the feedback inherent in an awarded grade. Grades are the symbolic representation of assessment outcomes and can provide global feedback to students regarding performance. But, as Wiggins (1998) emphasizes, students must not only receive feedback after their performance, but should also be given feedback during the performance so that they might self-adjust and improve. In fact, argues Wiggins, such self-adjustment (as opposed to self-assessment) should be the goal of any instruction. And the most effective way to promote self-adjustment is to provide feedback during a performance and then provide an opportunity for the student to apply the feedback. Post-performance global feedback is useful to a point, but specific and value-free responses to ongoing student performance are the keys to providing students with the information they need to truly improve. If the instructional goal is student self-adjustment, then the formative assessments during student performance are much more important than summative assessments given at the conclusion of the performance.

The second and fourth assumptions provide a compelling argument against grade inflation. If student evaluations are skewed toward the "A" range, then the ability of grades to discriminate between levels of performance is diminished because a smaller number of grade levels are used to describe the same number of students. But using a wider range of scoring levels is not necessarily useful unless students

are informed of the criteria used to discriminate between the levels. In the case of teacher preparation, national standards can provide specific indicators for discriminating between levels of performance.

The third assumption is a powerful statement regarding the rights of a learner. Too often, assessments are used solely to sort individuals by level of performance, a function that primarily aids the assessor and not the learner. Assessments must also serve the learner, and learners have the right to high-quality assessments that provide them with information that is useful to them in the process of learning.

The rationale behind this new grading policy is entirely consistent with our program's perspective on grading, but we experience a unique tension because we arguably provide more feedback during the practice teaching experience than does any other course at the university. During practice teaching we have the conscious goal of developing each student into an "A" level performer. Of course, we never completely attain that target, and the program strives to ensure the quality of the students we certify by using summative assessments as gatekeepers to the profession. We use convenient measures like GPA and PRAXIS scores to prevent low-performing students from entering the teaching force. We also use the Student Portfolio as a final check that students have tangible evidence of their teaching skill as defined by the New Jersey Professional Standards; students must pass the portfolio review before we submit their names for certification. Course grades in our program and letters of recommendation also provide information to potential employers regarding the quality of a student. However, the most effective way for our program to ensure the quality of our students is to put substantial resources into the formative evaluation of students and to provide feedback to students that will enable them to evolve professionally and to perform skillfully. Our staff meetings allow us to address, as a program, the strengths and weaknesses of each student and to use our collective wisdom for creating strategies to help each student develop their potential to the greatest degree.

When problems with student performance do arise, we can formulate responses that take into consideration as many variables and perspectives as possible, and we have confidence that our collective wisdom yields better strategies, decisions, and grades for our students than to rely on the cumulative score of a series of assessments or on

the professional judgment of a single person. By separating the process of grading from the process of coaching during the supervision process, we are able to establish productive relationships with our students that promote honest reflection on their teaching performance. Assigning grades to observations or evaluations would only confound the mentoring and coaching that takes place. Our challenge, though, is not to create a "black box" atmosphere where the determination of a grade is a mystery to our student. If we have done our job in providing detailed and honest feedback to students during their practice teaching and during the course of their portfolio development, then grades are never a surprise. Feedback and the reflection it generates with the student are the most important ingredients in the process of professional growth. If we can produce students who can reflect and then self-adjust, then we have taken a huge step towards producing beginning teachers who will evolve into master teachers. Thankfully, the university administration supports our efforts to produce as many "A" level teachers as possible, provided we submit convincing justification for each "A" awarded. We can only hope that our students find employment in professional settings that value the professional growth of teachers and provide feedback, and not just summative performance evaluation, that will promote their continued development as teachers.

Program Considerations and Evaluation

When New Jersey passed its new licensing code, the regulations contained a requirement that all programs of teacher preparation must be nationally accredited by 2009. Our program chose the Teacher Education Accreditation Council (TEAC) as the accrediting agency. TEAC requires programs in teacher preparation to use TEAC's three quality principles and standards for capacity to make "the case that its program has succeeded in preparing competent, caring, and qualified professional educators" (TEAC, 2005, ¶ 1). TEAC's three quality principles require programs to provide evidence of student learning, to demonstrate that the assessment of student learning is valid, and to demonstrate that the program exercises decisions based on evidence to improve program quality. The TEAC framework requires that each program develop a claims statement that describes the accomplishments of its students and graduates. The program faculty must then support

its claims statement with evidence that the claims have been realized. Finally, TEAC conducts an audit to verify the evidence submitted in support of the claims statement.

Our program claims that we prepare teachers who demonstrate teaching skill as defined by the New Jersey Professional Standards for Teachers. We will supply TEAC with ample evidence for that claim and its relationship to the TEAC quality principles from the assessments that we administer and from data that we collect about our students and program. The accreditation process is rigorous and demanding, to the point that we have formed a consortium of the New Jersey institutions that are being accredited by TEAC. More than half of the teacher preparation programs in the state are represented in this consortium, and the group meets with the purpose of aiding each other as we negotiate the challenges of accreditation. The TEAC accreditation process examines a wide range of evidence and emphasizes the analysis and interpretation of that evidence in terms of defined notions of quality. The process is designed to ultimately assist programs in using evidence and data to assess and improve the quality of their own work.

The rigor and complexity of the TEAC accreditation process serves as a stark contrast to the severely limited information provided by the Title II reporting and ranking process. Evaluation of a program and its ability to produce quality teachers is best left to accrediting agencies able to devote the time and resources for thoroughly evaluating a program with methodologies that reflect the complexities of the task. Government agencies can facilitate this process with policies that promote the transfer of information that programs and accrediting bodies find useful for program evaluation. For example, the real test of the quality of a program's preparation of students is how well its graduates perform as professionals. Tracking graduates, however, is an exceedingly difficult and time-consuming task. At the time of this writing, there is no mechanism in place within the State of New Jersey for providing programs in teacher preparation with employment data that would allow programs to follow their graduates into the teaching force. Locating graduates is only the first step—evaluating the performance of graduates once they take over their own classrooms is the next. Such evaluation would require access to pupil test data and to school-level teacher evaluation data.

Unfortunately, no mechanisms are in place to facilitate this flow of information from schools to teacher preparation programs, and most programs simply do not have the resources to seek out and collect the data themselves. Since these data already exist, the problem is primarily one of access. For institutions like Princeton, whose graduates often leave the state in which they are initially trained and certified, the problem is compounded with the movement between licensing jurisdictions. If government agencies could facilitate the flow of such information, programs would be better equipped to both follow and support their students during the first years of employment. Such steps might improve the quality of the nation's teaching force by allowing programs to monitor and mentor graduates as they enter the teaching force while also providing programs with valuable information on student professional performance that can be used to more fully inform program adjustment and improvement. The challenge for evaluating teaching at all stages is to be mindful that the collected data represents only a limited perspective of the complexity of teaching. Safeguards must accompany the collection and dissemination of such data to ensure that the information will be given sophisticated interpretation and will be used responsibly.

References

Malkiel, N. (2004, Summer). Explaining Princeton's new grading policy. *Princeton University Parents News, 27*(2), 9-11.

Teacher Education Accreditation Council. (2005). *TEAC accreditation process overview*. Retrieved March 18, 2005, from http://www.teac.org/accreditation/

Title II. (2005). *Title II frequently asked questions*. Retrieved February 15, 2005, from http://www.title2.org/faq.htm#rankings

Wiggins, G. (1998). *Educative assessment*. San Francisco: Jossey-Bass.

Todd W. Kent is the Associate Director of the Teacher Preparation Program at Princeton University.