From the Guest Editor:

Evaluating Teacher Education

Linda Darling-Hammond
Guest Editor
Stanford University

This volume of *Issues in Teacher Education* explores the possibilities and limitations of different approaches to assessing teacher learning and teacher education outcomes. It features five studies that illustrate a variety of methods for investigating the outcomes of teacher education; the studies also discuss the program implications of their findings.[1] In an era of increased attention to the outcomes of preparation programs (see, e.g., Cochran-Smith, 2001; Darling-Hammond, 2000b,c; Wise, 1996), teacher educators are working to develop strategies for assessing the results of their efforts—strategies that support their goals and conceptions of teaching, that appreciate the complexity of the teaching and teacher education enterprises, and that provide a variety of lenses on the process of learning to teach. The National Council for Accreditation of Teacher Education (NCATE) now requires that programs provide evidence of outcomes as they respond to each of the accreditation standards. In California, new regulations resulting from SB2042 require all teacher preparation programs to use a soon-to-be developed Teacher Performance Assessment (TPA) for evaluating their candidates or to develop their own version. Many programs are already developing a range of tools for gauging their candidates' abilities and their own success in adding to those abilities.

The goal of this collection is to contribute to the process of knowledge development about the assessment of teacher education that is already underway. We offer these studies not as evidence about the relative merit of a particular program, but as examples of what might be learned

from different kinds of studies of preparation and as grist for discussion of how findings can inform ongoing improvement. The studies examine different aspects of the work of the Stanford Teacher Education Program (STEP) using a variety of methods and evaluating a range of outcomes. These include:

- A written pre- and post-test assessment of foundational teaching knowledge, using the INTASC Test of Teaching Knowledge (TTK) as an instrument;
- An observational assessment of clinical teaching practice using a rubric based on the California Standards for the Teaching Profession (CSTP) with multiple samples of performance over time;
- A survey of the views of graduates about their preparedness for different dimensions of teaching and about their beliefs, practices, and career paths;
- An interview study of graduates' views of their learning in STEP, conducted with those who had prior experience in teaching before entering the program and who could, thus, reflect on what they learned from formal preparation vs. classroom experience alone;
- An interview and artifact analysis of students' learning with respect to the teaching of English Language learners, using syllabi and student work samples as well as interviews to evaluate what students had the opportunity to learn and what they did appear to learn.

Because the studies were undertaken during a period of reform of the program, some of them were intended to shed light on the outcomes of specific efforts to infuse new standards into the program, to redesign the curriculum, or to create stronger links between coursework and clinical work (see Hammerness & Darling-Hammond, this issue, for a discussion of these reforms). Other studies currently nearing completion (not included here) examine the teaching practices of graduates after they leave STEP (Hammerness, 2002), the results of surveys of employers, and the learning that results from the use of case writing in teacher education courses (Hammerness, Darling-Hammond, & Shulman, in press; Roeser, in press).

In developing these studies, we benefited from the work of many colleagues across the country who have struggled with these questions of how to assess teacher learning and teacher education through their work in individual teacher education programs and their work with

standard-setting bodies like the California Commission on Teacher Credentialing (CCTC), the Interstate New Teacher Assessment and Support Consortium (INTASC), and the National Board for Professional Teaching Standards (NBPTS). The questions and tools we brought to bear and the research strategies we used were greatly informed by their efforts.

Conceptualizing Outcomes of Teacher Education

Marilyn Cochran-Smith (2001) notes that:

> The question that is currently driving reform and policy in teacher education is what I refer to as "the outcomes question." This question asks how we should conceptualize and define the outcomes of teacher education for teacher learning, professional practice, and student learning . . . (p. 2)

Cochran-Smith identifies three ways that outcomes are currently being constructed: through evidence about the long-term impacts of teacher education on teaching practice and student learning; evidence about teacher test scores; and evidence about the professional performance of teacher candidates. Most of the studies described here deal with the third of these categories—outcome as "professional performance." However, the research reported on program graduates—through surveys, interviews, and observations—can be considered evidence of longer-term impacts of preparation on practice (the first category). The study of candidates' performance on the INTASC Test of Teaching Knowledge could be considered as an example of the second category, evidence about teacher test scores, although this within-program pre- and post-test measure of learning growth is a different use of tests than the aggregated results of standardized test scores required by Title II of the Higher Education Act, for example.

Developing professional performance is part of the core mission and daily work of teacher educators. How to assess the outcomes of this process is an issue that has been put front and center by the new outcomes-based emphasis of NCATE (Wise, 1996); the emphasis on teacher education outcomes in the federal Higher Education Act, which requires the evaluation of schools of education based on graduates' performance on standardized tests; and the growing policy debates about whether and how teacher education makes a difference to teacher effectiveness (see, e.g., Ballou & Podgursky, 2000; Darling-Hammond, 2000a). Cochran-Smith (2001) argues that a conception of standards is necessarily at the core of this approach to assessing teacher education: Constructing teacher education outcomes in terms of the professional

performances of teacher candidates begins with the premise that there is a professional knowledge base in teaching and teacher education based on general consensus about what it is that teachers and teacher candidates should know and be able to do. The obvious next step, then, is to ask how teacher educators will know when and if individual teacher candidates know and can do what they ought to know and be able to do. A related and larger issue is how evaluators (i.e. higher education institutions themselves, state departments of education, or national accrediting agencies) will know when and if teacher education programs and institutions are preparing teachers who know and can do what they ought to know and be able to do (p. 22).

Standards developed over the last decade by the NBPTS, the multistate INTASC consortium, NCATE, and California's CTC are closely aligned with one another and reflect a consensual, research-grounded view of what teachers should know and be able to do. Four of the five studies presented here define outcomes in ways that derive directly from these standards, and the fifth, which used open-ended interviews to record graduates' views of their own learning, resulted in categories that map onto the standards.[2]

The development of these studies occurred as the program was explicitly moving to integrate the CSTP and NBPTS standards into its curriculum and assessments for both coursework and clinical work. This standards integration process had the effect of clarifying goals, articulating for candidates the kinds of abilities they were expected to develop and for faculty and supervisors the kinds of supports and guidance they would need to provide. Thus, there was consonance between the program's efforts and the criteria against which candidate learning and program success were being evaluated. This consonance made the results of the studies much more usable in ongoing program reform than would have been the case if the measures of learning were out of synch with the program's intentions and aspirations.

The data represented in the studies include assessments of candidates' learning and performance from objective tests, from supervisors and cooperating teachers' observations, from work samples, from reports of candidates' practices, and from candidates' own perceptions of their preparedness and learning, both during the program and once they had begun teaching. We found analyses of these different sources of information intriguing, and, especially where the data could be triangulated across multiple data sources, productive for highlighting aspects of the program that were succeeding well or were in need of strengthening. Knowing that the results of teacher education are frequently only perceptible after candidates enter the classroom—sometimes years

later—we are also now eager to develop more research about candidates' performance, and their students' learning, while they are engaged in teaching.

What We Learned

We were interested in finding out what our candidates felt they had learned in the program (perceptual data collected through surveys and interviews); we also wanted to have independent measures of what they had learned (data from pre- and post-tests, work samples, and observations of practice over time). Finally, we wanted to know what our candidates did after they left STEP—whether they entered and stayed in teaching and what kinds of practices they engaged in (data from graduate surveys, which will soon be augmented with data from employers and direct observations of practice).

Strengths and Weaknesses

An obvious goal for evaluations of program outcomes is to identify areas where it appears the program is succeeding more and less well. Using different strategies allowed us to triangulate data from several sources to look for patterns in responses. Looking across several measures, we found, for example, confirmations that candidates felt well prepared in terms of planning and organizing curriculum in their subject matter and using a wide repertoire of teaching and assessment strategies adapted to student needs; that their supervisors saw substantial growth in these areas in terms of practice over the course of the year (Lotan & Marcus, this issue); and that test measures recorded growth in knowledge about these areas (Shultz, this issue). When compared to a national sample of beginning teachers, these were areas in which the program also appeared relatively strong (Darling-Hammond, Eiler, & Marcus, this issue).

We noted that areas in which the program appeared relatively strong compared to other programs were not always areas where we were fully satisfied. For example, even though 90 percent of STEP graduates reported feeling adequately prepared to teach English language learners (as compared to 50 percent of a national random sample of beginning teachers), fewer students felt "very well" prepared in this than in some other areas, and our more in-depth examination of the CLAD strand of courses and students' views (Bikle & Bunch, this issue) helped us to parse out which areas of their preparation were stronger (e.g. preparation to address diverse cultures and to support learning in the disciplines using

sheltered techniques) and which were weaker (e.g., preparation to teach English language skills to new English language learners).

We found some other areas where graduates felt less well-prepared. On our graduates' survey, fewer than 80 percent of graduates (proportions ranging from 73 to 79 percent) felt adequately prepared to identify and address special learning needs or difficulties, to work with parents, to use technology in the classroom, to create interdisciplinary curriculum, to resolve interpersonal conflict, and to assume leadership responsibilities in their school. Some of these are areas where teacher education programs have generally received lower ratings from their graduates (e.g. special education, technology use). Others are areas where our secondary program, heavily focused on content pedagogy, does less work than many elementary programs or those with a different orientation (e.g., creating interdisciplinary curriculum).

Making sense of these findings in program terms required triangulation with other data and an examination of trends over time (see below). These survey responses were sometimes reinforced by performance on the Test of Teaching Knowledge; for example, candidates' pre- and post-test score gains were partial in areas like responding to students' special needs, in which they showed increased understanding of the content requested in the question, but could not always discuss how they would apply their understanding to instructional practices (Shultz, this issue). We used these data to consider ongoing program reforms.

Effects of Program Reforms

One of the goals of the research was to uncover whether there were changes in candidates' learning over the three years that a number of program reforms were implemented (see Hammerness & Darling-Hammond, this issue, for a discussion of these changes). By collecting surveys from four years of program graduates we were able to examine whether there were changes in their views of certain aspects of the program over time. While there were not significant differences over time in most areas, there were some areas where program changes seemed to have made a large difference in graduates' feelings of preparedness. Some of these were positive and others were less so. On the one hand, the introduction of much more explicit work on how to use technology in the classroom, how to work with parents, and how to address special needs of exceptional students appeared to result in large increases in the proportions of graduates feeling adequately prepared in these domains (exceeding 80 percent in each category by 2000).

On the other hand, a sharp drop in candidates' felt readiness to

create interdisciplinary curriculum could also be attributed to program reforms. As efforts were made to tie courses more tightly together and streamline the curriculum to allow for the introduction of new content, a course that had earlier required an interdisciplinary curriculum project allowed students to use their discipline-based curriculum unit as the site for embedding required groupwork tasks. Thus, fewer students had the experience of constructing interdisciplinary curriculum. As in many program decisions, the faculty now needs to consider the trade-offs among competing goals for a one-year teacher education program and decide which values should guide a decision about whether or how to rethink the curriculum.

Another change—the infusion of CLAD as a core part of the program design—increased the exposure many students received to the knowledge and skill base needed to teach culturally and linguistically diverse students, but may have sacrificed some depth in the area of English Language Development. That, and the change in California outlawing bilingual education, put a course on Bilingual Education into an odd position in the curriculum. Data about student perceptions of preparedness allow the faculty to plan the ongoing redesign of this component in light of what students feel they know and can do and where they wish they knew still more.

Other Kinds of Outcomes

From a study of what already-experienced teachers felt they learned during this pre-service program, we learned some interesting things about the value that formal teacher education may add to the learning teachers feel they can get from experience alone (Kunzman, this issue). These teachers found, in particular, that they learned how to conceptualize and plan curriculum, recognize and work with struggling students, collaborate with other teachers, reflect productively on their practice in order to adjust and improve their plans, and use a theoretical framework for teaching as a way of making sense of classroom events. An analysis that tied this perceived learning back to specific courses and program experiences helped us to understand how some aspects of the program were working for these students. Discovering how much they valued certain kinds of learning opportunities encouraged us to maintain and expand certain components as we consider annual program changes. It has also clarified our thinking about how to educate already experienced teachers in a pre-service program—a phenomenon that is much more common in California than it is in other parts of the country.

In terms of other outcomes, we were interested to learn about the

career paths of our graduates and pleased to discover that almost all continued to hold teaching or other education positions, most in very diverse schools, and many had taken on leadership roles. Graduates who had been teaching longer reported that they felt more prepared to take on leadership roles. We suspect this is a function of experience more than preparation and hope to find out more about this in follow-up studies. We also want to pursue questions about the practices graduates engage in. While 80 percent or more reported engaging in practices we would view as compatible with the goals of the program, there was more variability in certain practices, such as using research to make decisions, involving students in goal-setting, and involving parents. We found that the use of these and other teaching practices is highly correlated with teachers' sense of preparedness. Teachers who felt most prepared were most likely to adjust teaching based on student progress and learning styles, to use research in making decisions, and to have students set some of their own learning goals and to assess their own work.

Equally interesting was the fact that graduates who felt better prepared were significantly more likely to feel highly efficacious—to believe they were making a difference and could have more effect on student learning than peers, home environment, or other factors. Although we found no relationship between the type of school a graduate taught in and the extent to which s/he felt efficacious or well-prepared, there are many important questions to be pursued about the extent to which practices and feelings of efficacy are related to aspects of the preparation experience and aspects of the teaching setting.

Using Data to Inform Program Changes

As noted above, these kinds of data can be used to fuel conversations about program changes and to examine the results of changes already made. We found it crucial to have several sources of data on the same question, including information that explicitly examines the connections between particular findings and specific aspects of the curriculum, in order to draw inferences about what is working well, what isn't, and what can be done about it. Not reported here are many nuances and details of the student feedback offered on specific course sections and sessions, supervisory groups, student teaching placements and student experiences that illuminated survey or interview findings or shed light on the results of the Test of Teaching Knowledge, the clinical observations, or student work samples. Without these, it would be much more difficult to draw inferences from the data that are useful for evaluating and developing appropriate changes.

Assessing Methods for Assessing Teacher Candidates and Programs

We learned much of interest about the possibilities and limits of different tools and strategies for evaluating teacher education candidates and program effects.

Surveys. For information about candidates' self-perceptions of preparedness across different dimensions of teaching, we found the use of a survey of graduates very helpful. Many programs use this kind of strategy for tracking graduates, particularly those that participate in NCATE reviews, which expect these kinds of data. Using a survey that was substantially derived from a national study of teacher education programs by the National Center for Restructuring Education, Schools, and Teaching (Darling-Hammond, 2000b) allowed us to compare our results to that of a national sample of beginning teachers. Conducting the survey with four cohorts and analyzing them separately and together allowed us to look at trends in graduates' perceptions of preparedness over time. Including data on practices was an important addition that we would expand upon in future in order to gain more information about what graduates report that they do in the classroom. We would also seek more information about the kinds of assignments and schools that characterize candidates' teaching jobs, so that we can understand in greater detail where our graduates go and how their experiences may interact with their practices and their feelings of preparedness.

We were very interested to learn in a factor analysis that graduates' responses to the survey loaded onto factors that closely mirror the California Standards for the Teaching Profession, a finding that suggests the validity of the standards as representing distinct and important dimensions of teaching.

Interviews of Students and Graduates. Interviews of students and graduates were an important adjunct to the survey findings, as they allowed us to triangulate findings and better understand the perceptions of candidates about how they were prepared. Candidates were asked not only about how prepared they felt but also about how they perceived the effects of specific courses and experiences. This explicit prompting allowed greater understanding of the relationships between program design decisions and student experiences. The interviewers were not instructors, and the responses were very candid and detailed. In one study, a course-by-course summary of graduates' comments prepared by the researcher provided excellent grist for re-evaluating specific courses.

In another, looking at interview data alongside samples of student work (which provided evidence of learning) and syllabi (which provided evidence of teaching), was extremely helpful in providing diagnostics that could inform program changes.

Other research has found that graduates' assessments of the utility of their teacher education experience evolve during their years in practice. With respect both to interviews and survey data, we would want to know how candidates who have been teaching for different amounts of time and in different contexts evaluate and re-evaluate what has been useful to them and what they wish they had learned in their pre-service program. Using survey data, it is not entirely possible to sort out these possible experience effects from the effects of program changes that affect cohorts differently. Interviews of graduates at different points in their careers that ask for such reflections about whether and when certain kinds of knowledge became meaningful for them would be needed to examine this more closely.

Also important is the collection of data on what candidates and graduates actually *do* in the classroom and what influences their decisions about practice. Whether it is possible to link such data on practices—which are connected to evidence about preparation—to evidence about relevant kinds of student learning is a question that many would like to answer, and one that is fraught with complexity. Examining the possibilities for developing these kinds of data are part of our plans for future research.

Longitudinal Observations of Clinical Practice. We learned several things about clinical assessment strategies from examining candidates' scores over time on a detailed rubric based on the CSTP standards. First, we learned that teacher candidates and supervisors viewed the rubric as very helpful in focusing their efforts and clarifying goals. Second, we learned from using the instrument in multiple observations that consensus between university supervisors and cooperating teachers (CTs) about the meaning of the rubric scores grew over time, either as a function of repeated use, conversations between supervisors and CTs, or the modest training efforts conducted by the program. The exact-score correlations between cooperating teachers' and supervisors' evaluations were very low at the beginning of the year and improved noticeably as the year went on. However, the correlations were never as high as would ideally be desirable, even if the assessments were generally very close. Thus, a third thing we learned is that the use of such assessments requires intensive, explicit efforts to develop shared meanings if they are to be viewed as reliable assessments for determining candidate recom-

mendations for certification and for conducting research on learning and performance. Finally, there are questions to be pursued regarding how one can independently confirm the improvements in practice that seem to be indicated by scores on an observational instrument through other measures of practice that could be used to validate these assessments.

Pre- and Post-Tests of Teaching Knowledge. A more unusual strategy for gauging learning was the use of the INTASC pilot Test of Teaching Knowledge (TTK) to look at pre- and post-program evidence about candidate knowledge of learning, development, teaching, and assessment. The TTK was developed by a group of teacher educators and state officials from the INTASC consortium, in collaboration with Educational Testing Service, to respond to the problem of teacher tests that have been critiqued for not testing teaching knowledge well—either because they focus only on basic skills or subject matter knowledge or because they ask questions about teaching in ways that are overly simplified, inauthentic, or merely require careful reading to discern the "right" answer (Haertel, 1991; Darling-Hammond, Wise, & Klein, 1999). For many years there have been press accounts of journalists and others not trained to teach who could take teacher competency tests and do as well as trained teachers because the content of the test so poorly represented the professional knowledge base. Whereas tests in some other professions are validated by comparing the scores of untrained novices with those of individuals who have received preparation (e.g., new law students vs. graduates of law school), this approach has not been used to validate teacher tests in the past.

Our experience with using the TTK at the beginning of the first quarter and end of the fourth quarter of a four-quarter preparation program was instructive in this regard. We were able both to document growth in learning for our candidates and provide evidence that, for the most part, the instrument appears to measure teaching knowledge that is acquired in a teacher education program. For most items, it was clear that most candidates knew very little at the start of their training and knew a great deal more (usually attaining the maximum score) at the end. However, seven of the twenty-six items appeared to suffer from some of the same flaws as items on earlier tests of teaching knowledge—that is, they were answerable by novices before they began their training because they required only a careful reading of the question or prompt to discern the desired response. In some cases, although the item appeared to be a valid measure of professional knowledge, the scoring rubric was designed in way that did not detect qualitative differences in responses. These findings suggest a need for further work on assessment development to enhance the validity of such measures.

Conclusion

Each of the kinds of tools we used has the potential to contribute different insights to an assessment of candidates' progress and program outcomes. Although each has limitations, we found them powerful in the aggregate for shedding light on the development of professional performance and how various program elements support this learning. We would like to develop even more powerful measures of performance—including means for evaluating the "teaching event" that candidates develop, videotape, and reflect upon as part of a culminating portfolio as well as systematic observations of graduates' practice—to supplement and validate these kinds of measures. Having examined a range of strategies, it seems to us that it will be important in this era of intense focus on single measures of teacher education outcomes to press for the use of multiple measures that allow a comprehensive view of what candidates learn and what a program contributes to their performance.

Notes

[1] The research reported in these papers was supported through generous grants from the William and Flora Hewlett Foundation and a Spencer Foundation Senior Scholars Grant.

[2] The examination of performance on the INTASC Test of Teaching Knowledge (Shultz) draws on the INTASC standards; the examination of candidate's clinical performance uses an observation instrument grounded in the CCTC's California Standards for the Teaching Profession (Lotan & Marcus); the survey of graduates examines dimensions of teaching drawn from the INTASC and National Board standards (Darling-Hammond, Eiler, & Marcus); the study of STEP graduates preparation to teach English language learners draws on the CCTC's CLAD standards (Bikle & Bunch). The study of graduates' views of their learning in STEP was open-ended, allowing graduates to identify the areas of learning they felt were important (Kunzman). Their definitions, however, map onto the standards as well, since they address core tasks of teaching well-represented in all of these standards.

References

Ballou, D. & Podgursky, M. (2000). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record*, *102* (1), 1-27.

Cochran-Smith, M. (2001). Constructing outcomes in teacher education: Policy, practice and pitfalls. *Education Policy Analysis Archives*, *9* (11), http://epaa.asu.edu/v9n1.

Darling-Hammond, L. (2000a). Reforming teacher preparation and licensing:

Debating the evidence. *Teachers College Record, 102* (1), 28-56.
Darling-Hammond, L. (2000b). *Studies of excellence in teacher education*. Washington, DC: American Association of Colleges for Teacher Education.
Darling-Hammond, L. (2000c). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8* (1). http://epaa.asu.edu/v8n1.
Darling-Hammond, L., Wise, A. E., & Klein, S. P. (1999). *A license to teach*. San Francisco: Jossey-Bass.
Haertel, E. H. (1991). New forms of teacher assessment. In G. Grant (Ed.), *Review of research in education, 17*, (pp. 3-29). Washington, DC: American Educational Research Association.
Hammerness, K. (2002, April). Looking for learning in practice: Examining the teaching practices of STEP graduates. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
Hammerness, K., Darling-Hammond, L., & Shulman, L. (in press). Toward expert thinking: How curriculum case-writing prompts the development of theory-based professional knowledge in student teachers. *Teaching Education*.
Roeser, R. (in press). The adolescent case: Bringing a 'whole-adolescent' perspective to secondary teacher education. *Teaching Education*.
Wise, A. E. (1996). Building a system of quality assurance for the teaching profession: Moving into the 21st century. *Phi Delta Kappan, 78* (3), 191-192.