Journal of Teacher Education

Assessing Teacher Education: The Usefulness of Multiple Measures for Assessing Program Outcomes

Linda Darling-Hammond

The online version of this article can be found at:
http://jte.sagepub.com/cgi/content/abstract/57/2/120

Additional services and information for *Journal of Teacher Education* can be found at:

Email Alerts: http://jte.sagepub.com/cgi/alerts

Subscriptions: http://jte.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

ASSESSING TEACHER EDUCATION
THE USEFULNESS OF MULTIPLE MEASURES
FOR ASSESSING PROGRAM OUTCOMES

Linda Darling-Hammond
Stanford University

Productive strategies for evaluating outcomes are becoming increasingly important for the improvement, and even the survival, of teacher education. This article describes a set of research and assessment strategies used to evaluate program outcomes in the Stanford Teacher Education Program during a period of program redesign over the past 5 years. These include perceptual data on what candidates feel they have learned in the program (through surveys and interviews) as well as independent measures of what they have learned (data from pretests and posttests, performance assessments, work samples, employers' surveys, and observations of practice). The article discusses the possibilities and limits of different tools for evaluating teachers and teacher education and describes future plans for assessing beginning teachers' performance in teacher education, their practices in the initial years of teaching, and their pupils' learning.

Keywords: *teacher education reform; teacher education*

Productive strategies for evaluating outcomes are becoming increasingly important for the improvement, and even the survival, of teacher education. In the political arena, debates about the legitimacy and utility of teacher education as an enterprise are being fought on the basis of presumptions—and some evidence—about whether and how preparation influences teachers' effectiveness, especially their ability to increase student learning in measurable ways (see, e.g., Darling-Hammond, 2000, in response to Ballou & Podgursky, 2000; Darling-Hammond & Youngs, 2002, in response to U.S. Department of Education, 2002). The federal Higher Education Act now requires that schools of education be evaluated based on graduates' performance on licensing tests, and the National Council for Accreditation of Teacher Education now requires that programs provide evidence of outcomes as they respond to each of the accredita-

tion standards (Wise, 1996). The Teachers for a New Era initiative launched by the Carnegie Corporation of New York and other foundations requires that the 11 institutions supported to redesign their programs collect evidence about how their teachers perform and how the students of these teachers achieve.

In light of these concerns, teacher educators are seeking to develop strategies for assessing the results of their efforts—strategies that appreciate the complexity of teaching and learning and that provide a variety of lenses on the process of learning to teach. Many programs are developing assessment tools for gauging their candidates' abilities and their own success as teacher educators in adding to those abilities. Commonly used measures range from candidate performance in courses, student teaching, and on various assessments used within programs to data on entry and retention in teach-

120

ing, as well as perceptions of preparedness on the part of candidates and their employers once they are in the field. In rare cases, programs have developed evidence of teachers' "impact" based on analyses of changes in their pupils' learning gauged through measures of student attitudes or behavior, work samples, performance assessments, or scores on standardized tests.

The impact or "effectiveness" data increasingly demanded by policy makers are, of course, the most difficult to collect and interpret for several reasons: First is the difficulty of developing or obtaining comparable premeasures and postmeasures of student learning that can gauge change in valid ways that educators feel appropriately reflect genuine learning; second is the difficulty of attributing changes in student attitudes or performances to an individual teacher, given all of the other factors influencing children, including other teachers past and present; third is the difficulty of attributing what the teacher knows or does to the influence of teacher education. Complex and costly research designs are needed to deal with these issues.

In this article, I describe a set of research and assessment strategies used to evaluate program outcomes in the Stanford Teacher Education Program (STEP) for the period of program redesign during the past 5 years, along with some of the findings from this research. In addition, I describe future plans for assessing beginning teachers' performance in teacher education, their practices in the initial years of teaching, and their pupils' learning. These plans include Stanford and a consortium of more than 15 California universities involved in the Performance Assessment for California Teachers (PACT) project, which has developed and validated a teacher performance assessment (TPA) used to examine the planning, instruction, assessment, and reflection skills of student teachers against professional standards of practice. We believe that these authentic assessments offer more valid measures of teaching knowledge and skill than traditional teacher tests, and they inspire useful changes in programs as they provide rich information about candidate abilities—goals

that are critical to an evaluation agenda that both documents and improves teacher education. Consortia of universities engaged in such assessments may also play a useful role in enabling the costly and difficult research on teacher effectiveness that policy makers desire. Finally, I discuss how these studies and tools have been and are being used to inform curriculum changes and program improvements.

BACKGROUND OF THE PROGRAM

The STEP program has historically been a 12-month postgraduate program in secondary education offering a master's degree and a California teaching credential.[1] Following a strongly critical evaluation conducted in 1998 (Fetterman et al., 1999), the program was substantially redesigned to address a range of concerns that are perennial in teacher education. These included a lack of common vision across the program; uneven quality of clinical placements and supervision; a fragmented curriculum with inconsistent faculty participation and inadequate attention to practical concerns such as classroom management, technology use, and literacy development; limited use of effective pedagogical strategies and modeling in courses; little articulation between courses and clinical work; and little connection between theory and practice (see also critiques of teacher education outlined in Goodlad, 1990; National Commission on Teaching and America's Future, 1996).

The STEP program traditionally also had several strengths. These included the involvement of senior faculty throughout the program, an emphasis on content pedagogy and on learning to teach reflectively, and a year-long clinical experience running in parallel with course work in the 1-year credential and master's degree program. The redesign of STEP sought to build on these strengths while implementing reforms based on a conceptual framework that infused a common vision that draws on professional teaching standards into course design, program assessments, and clinical work.

The program's conceptual framework is grounded in a view of teachers as reflective

practitioners and strategic decision makers who understand the processes of learning and development—including language acquisition and development—and who can use a wide repertoire of teaching strategies to enable diverse learners to master challenging content. A strong social justice orientation based on both commitment and skills for teaching diverse learners undergirds all aspects of the program. In addition to understanding learning and development in social and cultural contexts, professional knowledge bases include strong emphasis on content-specific pedagogical knowledge, literacy development across the curriculum, pedagogies for teaching special needs learners and English language learners, knowledge of how to develop and enact curriculum that includes ongoing formative and performance assessments, and skills for constructing and managing a purposeful classroom that incorporates skillful use of cooperative learning and student inquiry. Finally, candidates learn in a cohort and increasingly, in professional development school placements that create strong professional communities supporting skills for collaboration and leadership.

To create a more powerful program that would integrate theory and practice, faculty collaborated in redesigning courses to build on one another and add up to a coherent whole. Courses incorporated assignments and performance assessments (case studies of students, inquiries, analyses of teaching and learning, curriculum plans) to create concrete applications and connections to the year-long student teaching placement. Student teaching placements were overhauled to ensure that candidates would be placed with expert cooperating teachers (CTs) whose practice is compatible with the program's vision of good teaching. A "clinical curriculum" was developed on clearer expectations for what candidates would learn through carefully calibrated graduated responsibility and supervision on a detailed rubric articulating professional standards. Supervisors were trained in supervision strategies and the enactment of the standards-based evaluation system. In addition, technology uses were infused throughout the curriculum to ensure students' proficiency in integrating technology into their teaching.

Finally, the program sought to develop strong relationships with a smaller number of placement schools that are committed to strong equity-oriented practice with diverse learners. These have included several comprehensive high schools involved in restructuring and curriculum reform and several new, small, reform-minded high schools in low-income, "minority" communities, some of which were started in collaboration with the program. The guiding idea is that if prospective teachers are to learn about practice *in* practice (Ball & Cohen, 1999), the work of universities and schools must be tightly integrated and mutually reinforcing.

The secondary program has served between 60 and 75 candidates each year in five content areas—math, English, history/social science, sciences, and foreign language. A new elementary program will graduate about 25 candidates each year. During the course of the redesign, with enhanced recruitment, the diversity of the student body grew substantially, increasing from 15% to approximately 50% students of color in both the secondary and elementary cohorts.

It is clear that small programs like this one do not provide staff for large numbers of classrooms. Instead, they can play a special role in developing leaders for the profession if they can develop teachers who have sophisticated knowledge of teaching and are prepared not only to practice effectively in the classroom but also to take into account the "bigger picture" of schools and schooling—to both engage in state-of-the-art teaching and to be agents of change in their school communities. Indeed, in the San Francisco Bay Area, striking numbers of STEP graduates lead innovations and reforms as teachers, department chairpersons, school principals, school reform activists within and across schools, founders and leaders of special programs serving minority and low-income students, and increasingly, as new school founders. Thus, these leadership goals are explicit as part of the program's design for training. Described here are some of the studies and assessment tools thus far developed to evaluate how well these efforts are implemented and what the out-

comes are for preparedness, practice, and effectiveness in supporting student learning.

CONCEPTUALIZING OUTCOMES OF TEACHER EDUCATION

Assessing outcomes requires, first, a definition of what we expect teacher education to accomplish and influence in terms of candidate knowledge, skills, and dispositions and, second, means for measuring these things. As Marilyn Cochran-Smith (2001) has observed,

> 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 identified three ways that outcomes of teacher education are currently being considered:

1. through evidence about the professional performance of teacher candidates;
2. through evidence about teacher test scores; and
3. through evidence about impacts on teaching practice and student learning.

In what follows, I describe studies in each of these categories that seek to evaluate the candidate *learning* that occurs through particular courses and pedagogies, as well as through the program as a whole; the teaching *performance* of individuals as preservice candidates and as novice teachers; and the *outcomes* of this performance for students. With respect to the learning of students taught by STEP candidates, I describe the use of student learning evidence collected in the PACT teaching portfolio as a means for evaluating candidates' planning, instructional, and assessment abilities, and I describe a planned study that will examine evidence of student learning derived from standardized tests and performance assessments for students of beginning teachers who are graduates of STEP and other institutions. In addition, I describe the ways in which these studies and the assessment tools they have produced are used for ongoing *program improvement*, including changes in curriculum, pedagogy, and clinical supports.

Although we have conducted studies in all three of these categories, it is worth noting that most of the work falls in the first category—evidence about the professional performance of candidates. In this category, we include performance on teacher education assignments requiring analyses of teaching and learning—including a performance test of teacher knowledge (spilling over a bit into the second category)—as well as performance in the classroom during student teaching and (spilling into the third category) practices in the classroom during the 1st year of teaching. In all of these assessments, we agree with Cochran-Smith (2001) that a conception of standards is needed to productively examine teacher performance:

> 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)

This question is easier to address than it once was because of the performance-based standards developed during the past decade by the National Board for Professional Teaching Standards and the Interstate New Teacher Assessment and Support Consortium (INTASC), which has developed standards for beginning teacher licensing that have been adopted or adapted in more than 30 states. These have been integrated into the accreditation standards of the National Council for Accreditation of Teacher Education and reflect a consensual, research-grounded view of what teachers should know and be able to do. The studies presented here define outcomes related to candidates' knowledge and practice in ways that derive directly from these standards. Several use assessments developed on the standards (e.g., the INTASC test of teacher knowledge, a rubric used by su-

pervisors for evaluating student teaching performance based on the California Standards for the Teaching Profession—derived in turn from the INTASC standards, and a survey of program graduates developed to represent the dimensions of teaching included in the standards of the National Board for Professional Teaching Standards and INTASC).

The development of these studies occurred as the teacher education program was explicitly moving to integrate these standards into its curriculum and assessments for both course work 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. This created consonance between the program's efforts and the criteria against which candidate learning was being evaluated, and it made the results of the studies much more useful than would have been the case if measures of learning were out of sync with the program's aspirations.

The data represented in the studies include assessments of candidates' learning and performance from objective tests, from supervisors and CTs' observations in student teaching and from researchers' observations in the early years of teaching, 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. The PACT performance assessment allows systematic analysis of candidates' performances across different domains of teaching and comparison with those of other California teacher education programs. That assessment and the consortium of institutions involved in developing the assessment will enable future studies (also described below) that examine the effectiveness of teachers in terms of their students' learning gains in their 1st year of teaching.

TRACKING CANDIDATES' LEARNING

To examine what candidates learn in the STEP program, we have collected perceptual data on what they feel they have learned in the program (through surveys and interviews) as well as independent measures of what they have learned (data from pretests and posttests, performance assessments, work samples, and observations of practice). Finally, to learn about what our candidates do after they have left STEP—whether they enter and stay in teaching and what kinds of practices they engage in—we have used data from graduate surveys augmented with data from employers and direct observations of practice. We have learned much about the possibilities and limits of different tools and strategies for evaluating teacher education candidates and program effects.

Perceptual Data About Candidate Learning

Surveys. We developed a survey of graduates that has now been used for six cohorts of graduates to track perceptions of preparedness across multiple dimensions of teaching and provide data about beliefs and practices and information about career paths. Although there are limitations to self-report data—in particular the fact that candidates' feelings of preparedness may not reflect their actual practices or their success with students—research finds significant correlations between these perceptions and teachers' sense of self-efficacy (itself correlated with student achievement) as well as their retention in teaching (for a discussion, see Darling-Hammond, Chung, & Frelow, 2002). To triangulate these data, a companion survey of employers collects information about how well prepared principals and superintendents believe our graduates are along those same dimensions in comparison to others they hire. The survey was substantially derived from a national study of teacher education programs by the National Center for Restructuring Education, Schools, and Teaching (Darling-Hammond, in press), which allowed us to compare our results on many items to that of a national sample of beginning teachers.[2] Conducting the survey with four cohorts in the first round of research also allowed us to look at trends in graduates' perceptions of preparedness with time (Darling-

Hammond, Eiler, & Marcus, 2002) and to examine how our redesign efforts were changing those perceptions.

We learned 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 survey in representing distinct and important dimensions of teaching (see appendix.) We were pleased to discover that employers felt very positively about the skills of STEP graduates: On all of the dimensions of teaching measured, employers' ratings were above 4 on a 5-point scale, and 97% of employers gave the program the top rating of 5 on the question, "Overall, how well do you feel STEP prepares teacher candidates?" Of the employers, 100% said they were likely to hire STEP graduates in the future, offering comments such as, "STEP graduates are so well prepared that they have a huge advantage over virtually all other candidates," and "I'd hire a STEP graduate in a minute. . . . They are well prepared and generally accept broad responsibilities in the overall programs of a school." Program strengths frequently listed include strong academic and research training for teaching, repertoire of teaching skills and commitment to diverse learners, and preparation for leadership and school reform. Employers were less critical of candidates' preparedness than were candidates themselves, a finding similar to that of another study of several teacher education programs (Darling-Hammond, in press).

We were also pleased to learn that 87% of our graduates continued to hold teaching or other education positions, most in very diverse schools, and that many had taken on leadership roles. Most useful to us were data showing graduates' differential feelings of preparedness along different dimensions of teaching, which were directly useful in shaping ongoing reforms. However, given the limits of self-report data, these needed to be combined with other sources of data, as discussed in the Using Data for Program Improvement section below.

We also want to know about the practices graduates engage in. Although 80% or more reported engaging in practices we would view as compatible with the goals of the program, there was noticeable 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 was 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 assess their own work. Obvious questions arise about whether differences in the course sections to which candidates were assigned are related to these different practices.

Equally interesting is the fact that graduates who feel better prepared are significantly more likely to feel highly efficacious—to believe they are making a difference and can 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 she or he reported feeling 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.

Other research finds that graduates' assessments of the utility of their teacher education experiences 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 reevaluate what has been useful to them and what they wish they had learned in their preservice program. Using survey data, it is not entirely possible to sort out these possible experience effects from those 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 is examined further below.

Interviews of students and graduates. Interviews of students and graduates have been an important adjunct to survey findings, as they have allowed us to triangulate findings and better understand the perceptions of candidates about how they were prepared. We have used interviews in a number of studies and highlight three of them here as distinctive examples of how they have been helpful. In one instance, we explored the results of a particular course that had been redesigned; in another, a strand of courses was evaluated; and in a third, the effects of the program as a whole were examined. In all of these studies, candidates were asked not only about not only how prepared they felt but also about how they perceived the effects of specific courses and experiences. This explicit prompting—in conjunction with other data—allowed greater understanding of the relationships between program design decisions and student experiences.

In a study discussed by Roeser (2002), an instructor who had struggled with a course on adolescent development found that student evaluations improved significantly after the course was redesigned to include the introduction of an adolescent case study that linked all of the readings and class discussions into a clinical inquiry. The instructor conducted structured follow-up interviews with students after the conclusion of the course to examine their views of the learning experience as well as of adolescent students' development. He placed candidates' views of adolescent students in the context of a developmental trajectory of student teachers, documenting changes in their perspectives about adolescents as well as about their own roles and as teachers. These reports of candidate perspectives on their students, combined with their reports of their own learning and the data from confidential course evaluations collected with time, provided a rich set of information on what candidates learned and

what learning experiences were important to them.

In another study, researchers looked at learning in the Crosscultural, Language and Academic Development (CLAD) strand of courses and experiences intended to prepare candidates to teach culturally and linguistically diverse students (Bikle & Bunch, 2002). At the end of the year, the researchers conducted hour-long interviews with a set of students—selected to represent diverse subject areas and teaching placements—to understand how they felt their courses addressed the three domains of CLAD: (a) language structure and first and second language development; (b) methods of bilingual, English language development and content instruction; and (c) culture and cultural diversity. They reviewed course syllabi from eight courses that treated aspects of cultural and linguistic diversity to assess what instructors intended for students to learn in terms of these domains, and they reviewed student teachers' capstone portfolios to examine the extent to which candidates integrated course work and clinical experiences regarding the needs of English language learners into specific portfolio assignments.

The interviews not only explored what candidates learned in classes and applied to their placements but also placed this learning in the context of previous life experiences and future plans. Researchers asked for specific instances in courses and student teaching in which participants were able to connect classroom learning to practice or conversely, felt unprepared to deal with an issue of linguistic diversity. Finally, they asked candidates what would excite or concern them about teaching a large number of linguistically diverse students. The use of interview data—alongside samples of work from candidates' portfolios and syllabi—was extremely helpful in providing diagnostics that informed later program changes (discussed below.)

A third study examines what already-experienced teachers felt they learned during this preservice program (Kunzman, 2002, 2003), providing insights about the value that formal teacher education may add to the learning teachers feel they can get from experience alone.

About 20% of STEP students have already had at least 1 year of teaching experience before entering the preservice program. Unlike some programs serving teachers with experience, these teachers are fully integrated into the cohort, taking all the same courses and engaging in a full year of supervised student teaching like other candidates. Using a semistructured protocol, the author interviewed 23 of these STEP graduates from 1999 and 2000, asking them about their teaching experience prior to STEP and any training they might have had, their year of STEP study, and for 1999 graduates, their 1st year back in their own classroom since graduation.

Five themes emerged from interviews as areas of important learning for these experienced teachers: (a) increased effectiveness working with struggling students; (b) greater sophistication in curriculum planning, particularly in identifying and matching long-term objectives and assessment; (c) greater appreciation for collaborative teaching and ability to nurture collegial support; (d) structured opportunities for feedback and reflection on teaching practice; and (e) development of theoretical frameworks to support teaching skills and vision.

An analysis that tied this perceived learning back to specific courses and program experiences helped us to understand how various 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 considered annual program changes. The study also confirmed some of our decisions about how to educate already-experienced teachers in a preservice program—a phenomenon that is common in California where many individuals enter teaching without initial training. We concluded that these recruits appear to benefit at least as much as other candidates (in some cases perhaps more) from traditional student teaching in the classroom of an expert veteran and from a systematic set of courses that provides a conceptual framework and research base that both connects and corrects parts of their prior knowledge.

Analyses of Candidate Performance

Pretests and posttests 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 preprogram and postprogram evidence about candidate knowledge of learning, development, teaching, and assessment. The TTK was developed on the INTASC standards by a group of teacher educators and state officials from the INTASC consortium, in collaboration with Educational Testing Services. It aimed to respond to the problem of teacher tests that have been critiqued for not testing teaching knowledge well—either because they focus on only 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 (Darling-Hammond, Wise, & Klein, 1999; Haertel, 1991). 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 (Shultz, 2002). The 26 constructed response items on the pilot test we used are distributed across four sections. In the first section, candidates respond to 4 multiple-part questions addressing specific knowledge about learners and how that knowledge might influence the learning and/or teaching process. The second section asks candidates to read a case study or classroom

vignette focusing on aspects of learning, student behavior, or classroom instruction and to answer 7 questions related to the case study. The third section provides a "folio" or a collection of documents and asks candidates to answer 7 questions dealing with a particular learner or aspect of learning or teaching illustrated in the documents. In the final section, candidates answer 8 short, focused questions assessing propositional knowledge about specific theories, learning needs, instructional strategies, or teaching concepts.

For most items, it was clear that most candidates knew very little at the start of their training—in the pretest, candidates often wrote "I have no idea" or "I'm looking forward to learning about this during my year at STEP"—and they knew a great deal more (usually attaining the maximum score) at the end. However, 7 of the 26 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 a way that did not detect qualitative differences in responses. These findings suggest both the value of the test and a need for further refinement to enhance the validity of such measures.

Samples of student work. We studied how students learn to analyze their teaching by analyzing the several drafts of a curriculum case study they wrote in a course on principles of learning for teaching. In this course, case writing is designed to promote the application of learning theory to practical experiences in the classroom; a student-written curriculum case analyzing an instance of the candidate's own teaching serves as the central product of the class. The case focuses on the teaching of a curriculum segment with specific disciplinary goals, so that students will address central questions concerned with engaging students in the learning of subject matter. Students are asked to write about an incident in which they were trying to teach a key concept, problem, topic, or issue that is central

to the discipline, such as the concept of irony in English, evolution in science, pi in math, or the cultural differences in a foreign language. The incident may have been particularly successful, unsuccessful, surprising, or revealing and should have the potential to serve as a site for exploring interesting dilemmas or questions about teaching and learning. Student teachers must provide evidence of student learning to analyze how that learning (or lack of learning) was shaped by classroom decisions. (For a description of the process of developing this pedagogy, see Shulman, 1996.)

We examined data including students' cases (from outline to final draft), students' final self-assessment essays, interviews with instructors, and interviews with a sample of students (Hammerness, Darling-Hammond, & Shulman, 2002). Using the framework of "novice/expert" thinking proposed by Berliner (1986, 1991), we coded and scored student work, finding that students' successive case drafts demonstrated a development from naïve generalizations to sophisticated, theory-based explanations of the issues at play in their cases, characteristic of more "expert" thinking about teaching. We also found that certain aspects of the course pedagogy were important in helping student teachers learn to think like a teacher, including reading theoretical works in conjunction with writing cases; sharing cases with peer readers; receiving specific, theoretically grounded, concrete feedback from instructors; and revising the case several times in response to feedback about important elements of the context and teaching as well as potential theoretical explanations for what occurred.

Longitudinal observations of clinical practice. Another tool we developed to track candidates' learning is a detailed rubric for supervisors to use in evaluating student teaching progress, based on the California Standards for the Teaching Profession. This tool was informed by efforts at other institutions, especially the University of California campuses at Santa Barbara and Santa Cruz. Previous Stanford observation forms were entirely open-ended and produced widely differing kinds of observations of very different elements of teaching, de-

pending on what different observers thought to comment on. Research on assessment suggests that clear criteria are important for developing performance and that the usefulness of clinical experiences is weakened by lack of distinction between outstanding and ineffective teaching in assessment processes (Diamonti, 1977; McIntyre, Byrd, & Foxx, 1996), inadequate formative assessment (Howey & Zimpher, 1989), and a lack of clear roles for many supervisors and CTs (Cole & Knowles, 1995; Williams, Ramanathan, Smith, Cruz, & Lipsett, 1997).

Having specific indicators of each of the six California Standards for the Teaching Profession standards (the standards are noted in the appendix) and their associated substandards outlined on a scale from novice to expert provided guidance to supervisors and CTs in what to focus on (clarifying the content standards for clinical practice) and how to make judgments of performance—what counts as proficient performance adequate to sustain a recommendation for credentialing.

The relationship between these measures of performance in student teaching and what teachers do in "real" teaching is likely to depend in part on the nature and duration of the clinical experience. In this program, with a year-long student teaching placement, it is possible for candidates to gradually take on nearly all of the full responsibilities of a teacher, typically engaging in independent practice by February or March of the school year after assisting and coteaching for the 5 or 6 previous months. This allows teaching to be assessed as both a measure of candidate learning-in-progress and by the end of the year, as a proximal "outcome" of the overall preparation process. Furthermore, both the standards-based assessment instrument and to an even greater degree, the PACT assessment (described below) help to structure the kinds of performances candidates must engage in if they are to be assessed, thus, *creating* more systematic opportunities to learn and perform for student teachers than might otherwise occur by chance, given different contexts and expectations held by CTs.

We learned several things about clinical assessment strategies from examining candi-

dates' scores on this instrument: First, teacher candidates and supervisors viewed the rubric as helpful in focusing their efforts and clarifying goals. Second, we learned from using the instrument in multiple observations that consensus between university supervisors and CTs about the meaning of the rubric scores grew with time, probably as a function of repeated use, conversations between supervisors and CTs, and perhaps, the modest training efforts conducted by the program. The exact-score correlations between CTs' 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 recommendations for certification and for conducting research on learning and performance. Finally, there are questions about 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. I turn to these next.

ANALYZING PRACTICE AS AN OUTCOME OF PREPARATION

Although it is very helpful to look at candidates' learning in courses and their views of what they have learned, it is critical to examine whether and how they can apply what they have learned in the classroom. The problem of "enacting" knowledge in practice (Kennedy, 1999) is shared by all professions, but the problem is particularly difficult in teaching, where professionals must deal with large numbers of clients at one time, draw on many disparate kinds of knowledge and skill, balance competing goals, and put into action what they have learned while evaluating what is working from moment to moment and changing course as needed. To begin to explore whether our candidates can enact their learning in the classroom, we conducted two kinds of studies to examine

candidates' actual performance as teachers, both in the independent portion of the year-long student teaching they undertake as preservice candidates and as beginning teachers after they have graduated.

Observations of graduates' teaching practice. Hammerness (in press) first recorded program intentions through close analysis of syllabi and program documents and through interviews with faculty members; she then observed and interviewed 10 novice teacher graduates of the program using an observation form recording evidence of five key program elements in the graduates' practices. These elements include concern for students as learners and for their prior experiences and learning, the use of pedagogical content strategies to make subject matter accessible to students, commitment to equity, capacity to reflect, and commitment to change. Teachers' practice was coded as to whether there was "strong evidence," "some evidence," or "little evidence" of practice reflecting the 27 indicators of these elements.

The Hammerness (in press) study found that efforts to create program coherence on a set of themes were generally reflected in strong evidence of practices related to these themes. In particular, attention to students' needs and learning, use of well-grounded content pedagogical strategies, and commitment to equity for students were in strong evidence in virtually all of the graduates' practice. However, candidates felt less sure about their assessment practices than their other instructional approaches, and evidence of reflection and engagement in school change was spottier. These were areas identified for further curriculum work. Because this study included a careful analysis of syllabi across the program, as well as detailed observations of graduates' practices, it could inform specific changes in the curriculum (discussed below).

The PACT teaching assessment. Finally, the PACT assessment developed by a set of California universities has provided a means to evaluate elements of teaching skill systematically and authentically within the program. When California passed a law requiring a teacher per-

formance assessment (TPA) as a basis for programs' credentialing recommendations, the state developed its own TPA but gave colleges the option to develop their own and submit them, with evidence of validity and reliability, for approval. Twelve colleges created a consortium to develop a TPA—all of the University of California campuses, Stanford University and Mills College, plus 2 of the California State University campuses. This consortium has since grown to 17 programs and will continue to expand. The TPA created by the PACT consortium is modeled on both the National Board for Professional Teaching Standards' portfolio and on the portfolio for beginning teacher licensing used by the state of Connecticut.

The PACT includes a "teaching event" (TE) portfolio in the subject area(s) candidates teach plus "embedded signature assessments" used in each teacher education program (e.g., the development of curriculum units, child case studies, or analyses of learning). With modest philanthropic support and substantial in-kind contributions from the universities themselves, the assessments were piloted, scored, revised, and piloted again in academic years 2002-2003 and 2003-2004. During this period of time, more than 1,200 candidates at PACT institutions piloted TEs in the areas of elementary literacy and mathematics, English/language arts, history/social science, mathematics, and science. More than 250 teachers and teacher educators were trained to score these assessments in spring 2003 and spring 2004. Technical studies of reliability and validity have been conducted on these data (see Pecheone & Chung, 2006 [this issue], for details.)

For each TE, candidates complete several entries that are integrated on a unit or segment of instruction of about 1 week in length. These entries include

1. a description of their teaching context, including students and content;
2. a set of lesson plans from the segment of instruction;
3. one or two videotapes of instruction during the unit (depending on the field);
4. samples of student work during the unit; and
5. written reflections on instruction and student learning during the unit.

This collection of teacher and student artifacts is based on a planning, instruction, assessment, and reflection model in which candidates use knowledge of students' skills and abilities—as well as knowledge of content and how best to teach it—in planning, implementing, and assessing instruction. The planning, instruction, assessment, and reflection model is distinct in its placement of student learning at the center of the assessment system. Although many clinical assessments of preservice candidates focus on teacher activities and behaviors, paying little attention to evidence about student outcomes, the PACT TEs focus on evidence of student learning of defined objectives—including the learning of English language learners and students with learning differences—and ask candidates to consider the extent to which these objectives were attained for all students and how to adapt instruction to improve student learning.

There are several ways in which the PACT emphasizes attention to pupil learning. First, in the design of the instructional unit, candidates must describe how they have planned their unit based on what they know about their students' prior knowledge and learning and explain how their plans accommodate the needs of the group and individuals, including English language learners and students with exceptional needs. Second, as part of their planning, teachers show how they will incorporate formative as well as summative assessments in the unit and how they will use what they learn from the assessments to guide their teaching. Third, teachers teach the unit and record reflections each day about the students' responses and evidence of learning; then they describe how they will respond to students' needs in the next day's lesson. (Student teachers report this is a particularly powerful aspect of their PACT experience.) Fourth, candidates are asked to provide commentary on the videotapes they submit of themselves teaching part of the unit. The guiding questions they answer in this task, as well as others, focus on what they have observed about student learning of both specific disciplinary content and skills and of academic language.

Finally, candidates collect all of the student work from one assessment during the unit and analyze it in terms of what the work shows about student learning and areas for further teaching for different groups of students. This work is included in the portfolio for scoring, along with the teacher candidate's analysis and feedback to students. This evidence allows analysis of the kind and quality of work asked of and produced by students, how it reflects state standards and is aligned to what was taught, how well it was supported instructionally, and how closely and thoughtfully the teacher candidate can evaluate the work to understand what different students have learned and to plan for future instruction.

The PACT assessments provide evidence of candidate performance on authentic tasks of teaching scored in systematic ways that have allowed the participating universities to evaluate overall candidate performance, the relative strength of different areas of preparation (e.g., STEP candidates do better on planning, instruction, and reflection than they do on assessment), and the performance of candidates in comparison to those at other California institutions, which provides a broader perspective on our work and its success. Figure 1 illustrates some of the data available from the PACT, suggesting, for example, that scores are highest and most consistent across institutions on the planning task and increasingly variable for instruction, assessment, reflection, and language development. In general, scores are lowest on the assessment task, suggesting an area for attention across institutions. As described below, the PACT assessments will also provide a linchpin in a broader study of candidate effectiveness that examines practice and student learning gains.
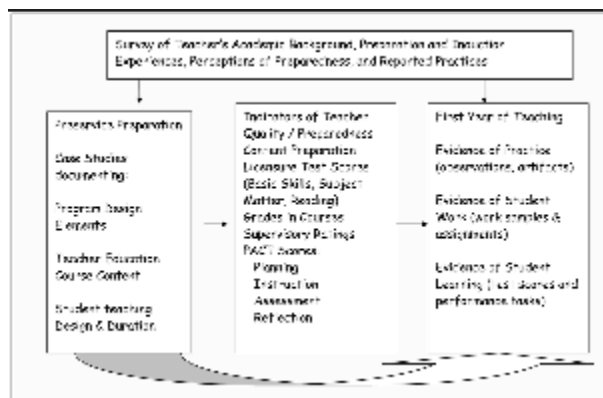
RESEARCH ON GRADUATES' EFFECTIVENESS

As noted earlier, the most difficult and, to many, the most important question is how what teachers have learned ultimately influences what their pupils learn. Even if teacher education students are followed into their classrooms,
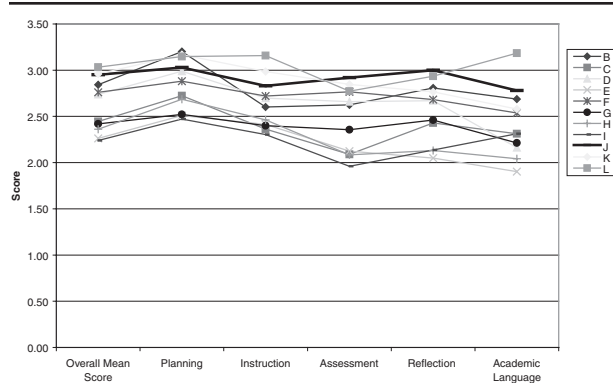
FIGURE 1: Performance Assessment for California Teachers Scores by Institution and Area, 2004

FIGURE 2: Data Collection Design
NOTE: PACT = Performance Assessment for California Teachers.

there are many complexities in approaching this question, including the problem of linking what teachers have learned to what they later do in the classroom—and then linking what they do to what their students learn, accounting for the variability in what these pupils bring with them. It is very difficult for most individual programs to be able to secure adequate data on these questions given the many and diverse districts and contexts their candidates leave to teach in, the small samples that can be tracked with any comparability, and the difficulty in securing useful and comparable pupil assessment data. We are seeking to approach this difficult question by capitalizing on the development of many of the assessments earlier described, including the PACT assessments, and by leveraging the cooperation of members of the PACT consortium to develop a large enough sample within a few large urban areas with enough variability in training to begin to link program features to practices and student outcomes.

The study will evaluate the practices and effectiveness of a sample of 300 to 400 elementary teacher education graduates from a number of the PACT universities, using measures of preservice teacher preparation experiences (documented components of programs and surveys measuring candidate perceptions of preparation and preparedness), preservice measures of teacher "quality" (e.g., grades, licensure test scores, supervisory ratings, and PACT scores in

literacy and mathematics), teacher practices in the classroom, and teacher effectiveness as evaluated by their students' achievement on both state standardized tests in literacy and mathematics and curriculum-based performance assessments given at the beginning and end of the school year that are more sensitive to higher order thinking and performance skills (see Figure 2.) These measures of student learning will allow analysis of both how students perform on large-scale assessments, controlling for their prior years' scores on these same tests, and how their performance has changed during the course of the school year on constructed response performance tasks that reflect the development of key reading, writing, and mathematics skills.

To build a chain of evidence, beginning teachers will be followed from the last year of their preparation into their 1st year of teaching. The analyses, using an approach rather like a path analysis, will evaluate the multiple connections among candidates' entering characteristics, their preparation experiences, performance as preservice candidates on traditional measures and the PACT performance assessment, and their practice as teachers. Teaching practice will be examined through observations and analysis of teaching artifacts such as lesson plans, videotapes, and student work samples. Even absent the consideration of pupil learning, these analyses will be valuable for exploring relationships among measures of performance and for begin-

ning to understand how what candidates encounter in their programs may influence what they are able to do in their classrooms.

It is clear that school contexts will have a large effect on teacher practice as well. Data on school contexts, including student demographics, working conditions, leadership and culture, and the nature of beginning teacher induction and supports, will be used to explore these relationships and to provide appropriate statistical controls. Multivariate multilevel analyses of the predictors of teacher effectiveness will be conducted, exploring the correlates of both practices and pupil learning gains including preservice components (e.g., course work elements, length and design of student teaching), other indicators of teacher quality (grades, test scores, background variables), indicators of teacher performance in preservice (supervisory ratings, PACT scores), and the amount and kind of induction support. While examining influences on teacher effectiveness, these analyses will also provide concurrent and predictive validity evidence about the PACT assessment as a measure of teacher quality.

The virtue of this design for examining teacher effectiveness is that in contrast to existing large-scale databases, the study will include more detailed measures of teacher education content and performance as well as broader measures of student achievement. It will treat teacher preparation and teaching as more than a black box. And in contrast to many small qualitative studies of individual programs, it will allow us to examine variation in preparation using both qualitative and quantitative measures of teacher performance and effects. However, even with these advantages, this approach will just begin to scratch the surface of the work to be done in establishing the relationships among aspects of preparation, teacher learning, teaching practice, and student learning that the field is wrestling with.

USING DATA FOR PROGRAM IMPROVEMENT

An obvious goal for evaluations of program outcomes is to identify areas where it appears the program is succeeding more and less well.

Another goal is to evaluate the effects of program reforms on candidates' opportunities to learn and on later performance. Using different strategies allowed us to triangulate data from several sources to look for patterns in responses.

Analyzing Strengths and Weaknesses

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 during the course of the year (Lotan & Marcus, 2002), and that test measures recorded growth in knowledge about these areas (Shultz, 2002). When compared to a national sample of beginning teachers, these were areas in which the program also appeared relatively strong (Darling-Hammond, Eiler, & Marcus, 2002).

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% of STEP graduates reported feeling adequately prepared to teach English language learners (as compared to 50% 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 students' experiences in the CLAD strand of courses helped us to parse out which areas of their preparation were stronger (e.g., preparation to address diverse cultures and to use "sheltered" techniques to teach content) and which were weaker (e.g., preparation to teach English language skills to new English language learners; Bikle & Bunch, 2002).

In addition, although most candidates felt well prepared to use a range of assessments, there was variation across subject matter areas; we observed less sophistication in the practice of some in this area, compared to areas such as planning and instruction, in both the follow-up observations of graduates and the PACT. This

confirms the importance of supplementing self-report instruments with other sources of data, and it points to an area of needed curriculum development. At the time of the survey, the program relied mostly on the subject-specific curriculum and instruction courses to teach about assessment, and they treated this topic very unevenly. This was addressed both by adding an assessment module in the practicum and asking faculty in subject-specific methods courses to discuss collectively what each was doing in the area of assessment and to define areas to develop further within those courses.

We found some other areas where graduates felt less well prepared. On our graduate survey, generally more than 80% of graduates felt adequately prepared for most of the tasks of teaching. However, somewhat smaller proportions (ranging from 73% to 79% when all 4 years of survey data were averaged) 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, such as creating interdisciplinary curriculum, are areas where our secondary program, which is heavily focused on content pedagogy within the disciplines, does less work than many elementary programs or those with a different orientation.

Making sense of these findings in program terms required triangulation with other data and an examination of trends with time (see below). These survey responses were sometimes reinforced by performance on the TTK. For example, candidates' pretest to posttest score gains were partial in areas such as 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, 2002). These findings led to the redesign of an instructional module on special education.

Analyzing the Effects of Program Reforms

One of the goals of the research was to uncover whether there were changes in candidates' learning during the 3 years that a number of program reforms were implemented (see Hammerness & Darling-Hammond, 2002, for a discussion of these changes). By collecting surveys from 4 years of program graduates, we were able to examine whether there were changes in their views of certain aspects of the program with time. Although there were not significant differences with 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 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% in each category by 2000).

On the other hand, a sharp drop in candidates' self-reported 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 group-work tasks. Thus, fewer students had the experience of constructing interdisciplinary curriculum. This project was reinstated. However, given the expectations for secondary school teachers in the field, their own felt needs, and the shortness of the program, we decided that other needs were more pressing than giving additional curriculum time to interdisciplinary curriculum.

As in many program decisions, the faculty needed to consider the trade-offs among competing goals for a 1-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 depth in the area of English language development. That and California's proposition outlawing bilingual education put a course on bilingual education into an odd position in the curriculum. Data about student perceptions of preparedness allowed the faculty to plan a redesign of this component in light of what students felt they knew and could do and where they wished they knew still more.

In using data to inform program changes, 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, to draw inferences about what is working well, what is not, and what can be done about it. More nuanced and detailed student feedback is also gathered from evaluations of specific course sections and sessions, supervisory groups, student teaching placements, and student experiences. These illuminate survey and interview findings and shed light on the results of the TTK, the clinical observations, and student work samples. Without course-specific information, it would be much more difficult to draw inferences from the data that are useful for evaluating and developing appropriate changes.

CONCLUSION

Each kind of tool described here has the potential to contribute different insights to an assessment of candidates' progress and program outcomes. Although each has limitations, we have 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 further refinement of the teaching event that candidates develop, videotape, and reflect on as part of a culminating portfolio, as well as more extensive systematic observations of graduates' practice and their students' outcomes—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.

APPENDIX
Teacher Survey Factors

Responses on a 5-Point Scale to "How well do you think your teacher preparation prepared you to . . ." (N = 152)	Loading Value	Item Mean (Standard Deviation)
Factor 1: Design curriculum and instruction (CSTP Standards 3, 4: Understanding and organizing subject matter for student learning; Planning instruction and designing learning experiences for all students)		
Q5: Develop curriculum that builds on students' experiences, interests and abilities.	0.720	4.02 (0.79)
Q7: Create interdisciplinary curriculum.	0.694	3.24 (1.13)
Q1: Teach the concepts, knowledge, and skills of your discipline(s) in ways that enable students to learn.	0.641	4.09 (0.80)
Q25: Use knowledge of learning, subject matter, curriculum, and student development to plan instruction.	0.582	4.10 (0.79)
Q9: Relate classroom learning to the real world.	0.581	3.72 (0.94)
Q14: Provide a rationale for teaching decisions to students/parents/colleagues.	0.500	3.93 (0.96)
Q6: Evaluate curriculum materials for their usefulness and appropriateness for your students.	0.484	3.76 (0.95)
Q18: Develop students' questioning and discussion skills.	0.457	3.68 (0.91)
Q8: Use instructional strategies that promote active student learning.	0.455	4.30 (0.68)
Factor 2: Support diverse learners (CSTP Standard 1: Engaging and supporting all students in learning)		
Q26: Understand how factors in the students' environment outside of school may influence their life and learning.	0.707	3.91 (0.90)
Q21: Teach students from a multicultural vantage point.	0.690	3.72 (.090)
Q24: Encourage students to see, question, and interpret ideas from diverse perspectives.	0.630	3.78 (0.86)
Q10: Understand how students' social, emotional, physical, and cognitive development influences learning.	0.552	3.95 (0.83)
Q19: Engage students in cooperative work as well as independent learning.	0.507	4.27 (0.75)
Q2: Understand how different students are learning.	0.472	3.97 (0.84)
Factor 3: Use assessment to guide learning and teaching (CSTP Standard 5: Assessing student learning)		
Q29: Give productive feedback to students to guide their learning.	0.669	3.68 (0.87)
Q30: Help students learn how to assess their own learning.	0.643	3.41 (0.82)
Q27: Work with parents and families to better understand students and to support their learning.	0.582	3.14 (0.86)
Q28: Use a variety of assessments (e.g., observation, portfolios, tests, performance tasks, anecdotal records) to determine student strengths, needs & programs.	0.488	4.09 (0.80)
Factor 4: Create a productive classroom environment (CSTP Standard 2: Creating and maintaining effective environments for student learning)		
Q34: Maintain discipline and an orderly, purposeful learning environment.	0.739	3.63 (0.93)
Q4: Help all students achieve high academic standards.	0.671	3.62 (0.83)
Q3: Set challenging and appropriate expectations of learning and performance for students.	0.603	3.85 (0.88)
Q15: Help students become self-motivated and self-directed.	0.482	3.43 (0.92)
Q12: Teach in ways that support new English language learners.	0.468	3.71 (1.05)
Q20: Use effective verbal and nonverbal communication strategies to guide student learning and behavior.	0.458	3.87 (0.89)
Factor 5: Develop professionally (CSTP Standard 6: Developing as a professional educator)		
Q36: Assume leadership responsibilities in your school.	0.700	3.32 (1.12)
Q35: Plan and solve problems with colleagues.	0.572	3.42 (1.06)
Q16: Use technology in the classroom.	0.566	3.21 (1.00)
Q33: Resolve interpersonal conflict.	0.528	3.14 (1.07)

NOTE: CSTP = California Standards for the Teaching Profession.

NOTES

1. In 2001, the program began a pathway for "co-term" students beginning in the undergraduate years; and in 2003, an elementary education program was added, beginning with Stanford undergraduates, who receive a disciplinary bachelor's degree and a master's in education in 5 years.

2. A very similar version of the survey was used in a study of 3,000 beginning teachers entering teaching through different programs and pathways in New York City. The results and the survey instrument are reported in Darling-Hammond, Chung, and Frelow (2002).

REFERENCES

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In. L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: A handbook of policy and practice* (pp. 3-32). San Francisco: Jossey-Bass.

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

Berliner, D. (1986). In pursuit of the expert pedagogue. *Educational Researcher*, *15*(7), 5-13.

Berliner, D. (1991). Perceptions of student behavior as a function of expertise. *Journal of Classroom Interaction*, *26*(1), 1-8.

Bikle, K., & Bunch, G. C. (2002). CLAD in STEP: One program's efforts to prepare teachers for linguistic and cultural diversity. *Issues in Teacher Education*, *11*(1), 85-98.

Cochran-Smith, M. (2001). Constructing outcomes in teacher education: Policy, practice and pitfalls. *Education Policy Analysis Archives*, *9*(11). Retrieved from http://epaa.asu.edu/epaa/v9n11.html

Cole, A. L., & Knowles, J. G. (1995). University supervisors and preservice teachers: Clarifying roles and negotiating relationships. *Teacher Educator*, *30*(3), 44-56.

Darling-Hammond, L. (2000). Reforming teacher preparation and licensing: Debating the evidence. *Teachers College Record*, *102*(1), 28-56.

Darling-Hammond, L. (in press). *Powerful teacher education: Lessons from exemplary programs*. San Francisco: Jossey-Bass.

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education*, *53*(4), 286-302.

Darling-Hammond, L., Eiler, M., & Marcus, A. (2002). Perceptions of preparation: Using survey data to assess teacher education outcomes. *Issues in Teacher Education*, *11*(1), 65-84.

Darling-Hammond, L., Wise, A. E., & Klein, S. P. (1999). *A license to teach*. San Francisco: Jossey-Bass.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers:" What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13-25.

Diamonti, M. C. (1977). Student teacher supervision. *Educational Forum*, *41*(4), 477-486.

Fetterman, D., Connors, W., Dunlap, K., Brower, G., Matos, T., & Paik, S. (1999). *Stanford Teacher Education Program 1997-98 evaluation report*. Stanford, CA: Stanford University.

Goodlad, J. I. (1990). *Teachers for our nation's schools*. San Francisco: Jossey-Bass.

Haertel, E. H. (1991). New forms of teacher assessment. In G. Grant (Ed.), *Review of research in education* (Vol. 17, pp. 3-29). Washington, DC: American Educational Research Association.

Hammerness, K. (in press). From coherence in theory to coherence in practice. *Teachers College Record*.

Hammerness, K., & Darling-Hammond, L. (2002). Meeting old challenges and new demands: The redesign of the Stanford Teacher Education Program. *Issues in Teacher Education*, *11*(1), 17-30.

Hammerness, K., Darling-Hammond, L., & Shulman, L. (2002). Toward expert thinking: How curriculum case writing prompts the development of theory-based professional knowledge in student teachers. *Teaching Education*, *13*(2), 221-245.

Howey, K. R., & Zimpher, N. L. (1989). *Profiles of preservice teacher education*. Albany: State University of New York Press.

Kennedy, M. (1999). The role of preservice teacher education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 54-85). San Francisco: Jossey Bass.

Kunzman, R. (2002). Preservice education for experienced teachers: What STEP teaches those who have already taught. *Issues in Teacher Education*, *11*(1), 99-112.

Kunzman, R. (2003). From teacher to student: The value of teacher education for experienced teachers. *Journal of Teacher Education*, *54*, 241-253.

Lotan, R., & Marcus, A. (2002). Standards-based assessment of teacher candidates' performance in clinical practice. *Issues in Teacher Education*, *11*(1), 31-48.

McIntyre, J. D., Byrd, D. M., & Foxx, S. M. (1996). Field and laboratory experiences. In J. Sikula, T. J. Buttery, & E. Guyton (Eds.), *Handbook of Research on Teacher Education* (2nd ed., pp. 171-193). New York: Macmillan.

National Commission on Teaching and America's Future. (1996). *What matters most: Teaching for America's future*. New York: Author.

Pecheone, R., & Chung, R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education*, *57*, 22-36.

Roeser, R. W. (2002). Bringing a "whole adolescent" perspective to secondary teacher education: A case study of the use of an adolescent case study. *Teaching Education*, *13*(2), 155-179.

Shulman, L. S. (1996). Just in case: Reflections on learning from experience. In J. Colbert, P. Desberg, & K. Trimble (Eds.), *The case for education: Contemporary approaches for using case methods* (pp. 197-217). Boston: Allyn & Bacon.

Shultz, S. (2002). Assessing growth in teacher knowledge. *Issues in Teacher Education*, *11*(1), 49-64.

U.S. Department of Education. (2002). *The secretary's report on teacher quality*. Washington, DC: Author.

Williams, D. A., Ramanathan, H., Smith, D., Cruz, J., & Lipsett, L. (1997). Problems related to participants' roles and programmatic goals in student teaching. *Mid-Western Educational Researcher*, *10*(4), 2-10.

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.

Linda Darling-Hammond *is Charles E. Ducommun Professor of Education at Stanford University where she was faculty sponsor of the Stanford Teacher Education Program from 1998 to 2004. Her research, teaching, and policy interests focus on teaching quality, school reform, and educational equity. She is coeditor of the National Academy of Education's recent volume* Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able To Do *(Jossey-Bass, 2005) and its companion volume,* A Good Teacher in Every Classroom: Preparing the Highly Qualified Teachers Our Children Deserve *(Jossey-Bass, 2005).*