Edith Cowan Peer Reviewed Journals & Books Publishers

Edith Cowan Journal of Education
ISSN: 2790-055X
Vol.2, Issue No.1, pp1-17.

Evaluating teacher education outcomes: A study of the Stanford teacher education programme.

[1]Abigail Jackson and [2]Charlotte Mia
[12]Department of Education, School of Education, Stanford University.
Corresponding author's e-mail: abigail.jackson@gmail.com

ARTICLE INFO

Article history:
Received Date: 2nd Nov 2021
Revised Date: 14th April 2022
Accepted Date: 18th May 2022

Keywords:
Teacher education; programme evaluation; student outcomes; valueadded assessment.

A B S T R A C T

This paper considers a set of research and assessment strategies used to evaluate program outcomes in the Stanford Teacher Education Program (STEP) during a period of program redesign over 10 years. These strategies include surveys and interviews of students' perceptions of program elements and their own preparedness, observations of their practice during and after teacher education, evaluations of their practice on a structured portfolio of practice (the Performance Assessment for California Teachers (PACT)) and analyses of the effects of a sample of graduates of STEP and other programmes on student outcomes, including value-added measures. While the studies were able to ascertain that the students of STEP graduates had strong value-added learning gains, the paper concludes that the use of student learning data alone as a measure of teacher effectiveness does not help guide decisions related to programme improvement, and a range of approaches is required. In addition, it suggests that there will be continuing concerns about the narrowness of the learning measured by standardized tests, and about the many challenges of collecting and analyzing such data in ways that overcome the technical and practical problems associated with their use.

Background

Productive strategies for evaluating teacher education outcomes are becoming increasingly important for the improvement, and even the survival, of the enterprise. In the political arena in a number of countries, debates about the utility of teacher preparation are being fought on the basis of evidence about whether and how it influences teachers' effectiveness, especially their ability to increase student learning in measurable ways (e.g. Darling-Hammond and Youngs 2002, in response to US Department of Education 2002).

In the USA, the Higher Education Act requires that schools of education be evaluated based on graduates' performance on licensing tests, and a new federal funding initiative, entitled 'Race to the Top', encourages states to create databases that link teachers to their students' test scores and to use these data to evaluate the effectiveness of both teacher education programmes and individual teachers. This continues a trend toward more outcome-based evaluation of programmes begun when the Teachers for a New Era initiative launched by the Carnegie Corporation of New York and other foundations required that the 11 institutions

supported to redesign their programmes existence of considerable supply management literature, there is little empirical investigation of the role of supply management in building supply flexibility for greater business performance. There are three specific gaps in the literature. First, while most companies recognize that flexibility is key to success, many have not collect evidence about how their teacher candidates 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 effort, strategies that appreciate the complexity of teaching and learning and that provide a variety of lenses on the process of learning to teach. Many programmes are developing assessment tools for gauging their candidates' abilities and their own success as teacher educators in adding to those abilities. Measures commonly used range from candidate performance in courses, student teaching, and on various assessments used within programmes to data on entry and retention in teaching, as well as perceptions of preparedness on the part of candidates and their employers once they are in the field. In rare cases, programmes have developed evidence of teachers' impact based on analyses of changes in their pupils' learning gauged through measures of student attitudes or behaviour, work samples, performance assessments, or scores on standardised tests.

In this article, we describe a set of research and assessment strategies used to evaluate programme outcomes in the Stanford Teacher Education Programme (STEP) during a period of programme redesign over the course of a decade, along with some of the findings from this research. These data include the usual surveys and interviews of student perceptions of programme elements and their own preparedness, along with observations of their practice during and after teacher education, evaluations of their practice on a structured portfolio of practice (the Performance Assessment for California Teachers (PACT) that is used to license beginning teachers in more than 30 institutions in California), and analyses of the effects of a sample of graduates of STEP and other programmes on student outcomes

This study is noteworthy in several ways. First, the trend to measure teacher effectiveness and teacher education programme quality with standardised tests is a controversial and ever more prominent part of the USA's federal reform agenda, also increasingly being discussed in other countries. The impact or 'effectiveness' data demanded by policymakers are the most difficult to collect and interpret for several reasons. First is the difficulty of developing or obtaining valid comparable pre- and post-measures of student learning change 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 research, a multi-faceted research agenda was underway for a number of years to examine the processes of teacher learning and a wide range of teacher education outcomes, allowing triangulation with the data on student achievement gains later collected. While there is reason to be extremely cautious in interpreting any such standardised test data, this triangulation with other findings allows somewhat greater confidence in the results of the analysis, and greater capacity to hypothesise about the nature of programme contributions to candidate capacities. Also unusual is the fact that the continuing STEP programme redesign process was informed by extensive data about candidates' and faculty learning, and by employers' perceptions of teachers prepared in the programme. Such evidence-driven reform efforts are uncommon in the USA, where most teacher education reform is driven by the relatively infrequent accreditation process,

which documents inputs to the educational process, rather than outcomes and effects of change efforts (Richardson and Roosevelt 2004).

Literature Review

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

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

The programme'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 programme. 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 formative and performance assessments, and skills for constructing and managing a purposeful classroom that incorporates skilful 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 programme that would better integrate theory and practice and allow candidates to be more successful with diverse learners in highneed schools and communities, a number of steps were taken: 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 whose practice is compatible with the programme's vision of good teaching; a 'clinical curriculum' was developed around clearer expectations for what candidates would learn through carefully calibrated graduated responsibility and supervision around a detailed rubric articulating professional standards; and 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 programme 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, reformminded high schools in low-income, 'minority' communities, some of which were started in collaboration with the programme. The guiding idea is that if prospective teachers are to learn about practice in practice (Ball and Cohen 1999), the work of universities and schools must be tightly integrated and mutually reinforcing.

The secondary programme has served between 60 and 75 candidates each year in five content areas (mathematics, English, history/social science, sciences, and a foreign language). A newer elementary programme graduates about 20–25 candidates each year. Over the course of the redesign, with enhanced recruitment, the racial/ ethnic diversity of the student body grew substantially, increasing from 15% to approximately 50% teacher candidates of colour in both the secondary and elementary cohorts.

Clearly, small programmes like this one do not provide staff for large numbers of classrooms. Instead, they can play a 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, so as both to 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 programmes serving minority and low-income students, and, increasingly, as new school founders. Thus, these leadership goals are explicit as part of the programme'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 outcomes are for preparedness, practice, and effectiveness in supporting student learning.

Materials and Methods

Perceptual data about candidate learning

Surveys

A quantitative survey has been used for repeated cohorts of graduates to track perceptions of preparedness across multiple dimensions of teaching, provide data about beliefs and practices, and information about career paths. While there are limitations of self-report data, in particular the fact that candidates' feelings of preparedness may not reflect their actual practices or their success with students, research has found significant correlations between these perceptions and teachers' sense of self-efficacy (itself correlated with student achievement) as well as their retention in teaching (see Darling-Hammond, Chung, and Frelow 2002). To triangulate these data, a companion survey of employers collected information about how well prepared principals and superintendents believe STEP graduates are in comparison to others they hire.

The survey was substantially derived from a national study of teacher education programmes, which allowed us to compare STEP results on many items to those from a national sample of beginning teachers (Darling-Hammond 2006a). Conducting the survey with four cohorts in the first round of research also allowed us to look at trends in graduates' perceptions

of preparedness over time (Darling-Hammond, Eiler, and Marcus 2002), and to examine how the programme redesign efforts were changing those perceptions.

A factor analysis revealed that graduates' responses to the survey loaded onto factors that closely mirror the California Standards for the Teaching Profession: Designing Curriculum and Instruction, Supporting Diverse Learners, Using Assessment to Guide Learning and Teaching, Creating a Productive Classroom Environment, and Developing Professionally. This finding suggests the validity of the survey in representing distinct and important dimensions of teaching (for further detail see Darling-Hammond 2006b).

The survey found 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 five-point scale, and 97% of employers gave the programme the top rating of '5' on the question: 'Overall, how well do you feel STEP prepares teacher candidates?' One-hundred per cent said they were likely to hire STEP graduates in the future, offering comments like, '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 programmes of a school.' Programme strengths frequently listed included 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 a set of several teacher education programmes (Darling-Hammond 2006a)

The survey also revealed that 87% of STEP 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 programme leaders were data showing graduates' differential feelings of preparedness along different dimensions of teaching, which were directly useful in shaping continuing reforms. However, given the limits of selfreport data, these needed to be combined with other sources of data, as discussed further below.

The survey asked about the practices graduates engage in. While 80% or more reported engaging in practices viewed as compatible with the goals of the programme, 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 to assess their own work. Obvious questions arise about whether these different practices are related to differences in the course sections or cooperating teachers to which candidates were assigned.

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 he or she reported feeling efficacious or wellprepared, 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 has found 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 re-evaluate what has been useful to them and what they wish they had learned in their pre-service programme. Using survey data, it is not entirely possible to sort out these possible experience effects from those of programme 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 teacher candidates and graduates and their learning

Interviews of teacher candidates and graduates are an important adjunct to survey findings, as they allow triangulation of findings and a better understanding of the perceptions of candidates about how well they were prepared. Three studies are highlighted here as distinctive examples of how interviews can be helpful. In one instance, researchers 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 programme as a whole were examined. In all of these studies, 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, in conjunction with other data, allowed greater understanding of the relationships between programme design decisions and student experiences.

In one study, 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 which linked all of the readings and class discussions into a clinical inquiry (Roeser 2002). He conducted structured follow up interviews with students after the course was over 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 over 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 strand of courses and experiences intended to prepare candidates to teach culturally and linguistically diverse students (Bikle and 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 so as to understand how they felt their courses addressed the three domains of California's requirements: (1) language structure and first and second language development; (2) methods of bilingual English language development and content instruction; and (3) 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 coursework and clinical experiences regarding the needs of English language learners into specific portfolio assignments.

The interviews explored not only 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 programme changes (see below).

A third study examined what experienced teachers felt they learned during this pre-service programme (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. At the time of the study, about 20% of STEP students had already had at least a year of teaching experience before entering the pre-service programme. Unlike some programmes 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 semi-structured protocol, the author interviewed 23 of these STEP graduates from two cohorts, asking them about their teaching experience prior to STEP and any training they might have had; their year of STEP study; and their first year back in their own classroom since graduation.

Five themes emerged from interviews as areas of important learning for these experienced teachers:

(1) increased effectiveness working with struggling students;

(2) greater sophistication in curriculum planning, particularly in identifying and matching long-term objectives and assessment;

(3) greater appreciation for collaborative teaching and ability to nurture collegial support;

(4) structured opportunities for feedback and reflection on teaching practice; and

(5) development of theoretical frameworks to support teaching skills and vision.

An analysis that tied this perceived learning back to specific courses and programme experiences helped reveal how various aspects of the programme were working for these candidates. Discovering how much they valued certain kinds of learning opportunities encouraged STEP leaders to maintain and expand certain components as annual programme changes were considered. The study also confirmed some decisions about how to educate already experienced teachers in a pre-service programme, a phenomenon that is common in California where many individuals enter teaching without initial training. The study confirmed 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 provide a conceptual framework and research base that both connects and corrects parts of their prior knowledge.

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-programme evidence about candidate knowledge of learning, development, teaching, and assessment. The TTK was developed around the INTASC standards by a group of teacher educators and state officials from the INTASC consortium, in collaboration with Educational Testing Service (ETS). It aimed 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, and 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.

STEP's review of its experience with using the TTK at the beginning of the first quarter and end of the fourth quarter of a four-quarter preparation programme was instructive in this regard. The study documented growth in learning for STEP candidates and provided evidence that the instrument appeared to measure teaching knowledge that is acquired in a teacher education programme (Shultz 2002). The 26 constructed response items on the pilot test were distributed across four sections. In the first section, candidates responded to four multiple part questions addressing specific knowledge about learners and how that knowledge might influence the learning and/or teaching process. The second section asked candidates to read a case study or classroom vignette focusing on aspects of learning, student behaviour, or classroom instruction and to answer seven questions related to the case study. The third section provided a 'folio' or a collection of documents and asked candidates to answer seven questions dealing with a particular learner or aspect of learning or teaching illustrated in the documents. In the final section, candidates answered eight 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 pre-test, candidates often wrote 'I have no idea,' or 'I'm looking forward to learning about this during my year at STEP.' They knew a great deal more at the end of the year, with a large majority attaining the maximum score on nearly all items. However, seven of the 27 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 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 both the value of the test and a need for further refinement to enhance the validity of such measures.

Samples of student work

Another study examined how students learn to analyses 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 was designed to promote the application of learning theory to practical experiences in the classroom; a student written curriculum case analyzing an instance of the candidates' own teaching serves as the central product of the class. The case focused on the teaching of a curriculum segment with specific disciplinary goals, so that students would address central questions concerned with engaging students in the learning of subject matter. Students were 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 mathematics, or the cultural differences in a foreign language. The incident might 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 provided evidence of student learning in order to analyse

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.)

A study examined candidates' cases (from outline to final draft), their final self-assessment essays, interviews with instructors, and interviews with a sample of students (Hammerness, Darling-Hammond, and Shulman 2002). Using the framework of 'novice/expert' thinking proposed by Berliner (1986, 1991), student work was coded and scored. The study concluded that students' successive case drafts demonstrated a development from naïve generalizations to sophisticated, theory-based explanations of the issues at play in the cases, characteristic of more 'expert' thinking about teaching. The analysis also revealed 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.

Analyses of candidate performance

Longitudinal observations of clinical practice

Another tool STEP developed to track candidates' performance, as well as their 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, depending 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, and Foxx 1996), inadequate formative assessment (Howey and Zimpher 1989), and a lack of clear roles for many supervisors and cooperating teachers (Williams et al. 1997; Cole and Knowles 1995).

Having specific indicators of each of the six CSTP standards (for further detail see Darling-Hammond 2006b) and their associated sub-standards outlined individually on a scale from novice to expert provided guidance to supervisors and cooperating teachers in what to focus on and observe (clarifying the content standards for clinical practice) and how to make judgements of performance regarding what counts as proficient performance adequate to sustain a recommendation for the award of the appropriate credential.

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 programme, 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 co-teaching for the five or six previous months. This allows teaching to be assessed as both a measure of candidate learningin-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 cooperating teacherss.

Candidates' scores over time on this instrument revealed several things. 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 cooperating teachers (CTs) about the meaning of the rubric scores grew over time, probably as a function of repeated use, conversations between supervisors and CTs, and, perhaps, the modest training efforts conducted by the programme. 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 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. We turn to these next.

Analysing practice as an outcome of preparation

While 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 one that 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 STEP candidates can enact their learning in the classroom, two kinds of studies were conducted to examine candidates' actual performance as teachers, both in the independent portion of the year-long student teaching they undertake as pre-service candidates and as beginning teachers after they have graduated.

Observations of graduates' teaching practice

One study documented programme intentions through close analysis of syllabi and programme documents and through interviews with faculty members, and then observed and interviewed 10 novice teacher graduates of the programme using an observation form that sought evidence of five key programme elements in the graduates' practices (Hammerness 2006). These elements included 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 research found that efforts to create programme coherence around 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 more spotty. These were areas identified for further curriculum work. Because this study included a careful analysis of syllabi across the programme, 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 programme. When California passed a law requiring a teacher performance assessment (TPA) as a basis for programmes' credentialing recommendations, the state developed its own TPA, but gave colleges the option to develop their own TPAs and submit them, with evidence of validity and reliability, for approval. Twelve colleges created a consortium to develop a teacher performance assessment (all of the University of California campuses, plus Stanford University and Mills College, plus two of the California State University campuses). This consortium has since grown to over 30 programmes and will continue to expand. The teacher performance assessment created by the PACT consortium is modelled in many respects on both the National Board for Professional Teaching Standards' portfolio and on the portfolio for beginning teacher licensing developed by the state of Connecticut.

The PACT includes a 'teaching event' portfolio in the subject area(s) candidates teach, plus 'embedded signature assessments' used in each teacher education programme (for example, 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 in 2002– 03, validated as technically valid and reliable, and are currently in use as tools for licensure recommendations (see Pecheone and Chung 2006).

For each teaching event (TE), candidates complete several entries that are integrated around a unit or segment of instruction of about one 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 artefacts is based on a Planning, Instruction, Assessment, and Reflection (PIAR) 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 model is distinct in its placement of student learning at the centre of the assessment system. While many clinical assessments of pre-service candidates focus on teacher activities and behaviours, paying little attention to evidence about student outcomes, the PACT Teaching Events focus on evidence of student learning of defined objectives, including the learning of English language learners and students with learning differences (students with formally diagnosed learning disabilities and others who learn in nontraditional ways), 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 s, 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 (and student teachers report this is a particularly powerful aspect of their PACT experience).

Fourth, candidates are asked to provide commentary of 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 and the performance of candidates in comparison to those at other California institutions, which provides a broader perspective on each programme's work, and opportunities for programmes to learn from one another. STEP candidates' scores on the PACT have compared favourably to the average scores of candidates at other institutions (see Figure 1).

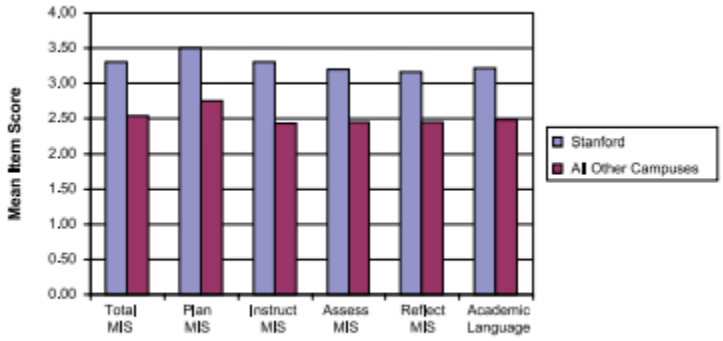

Figure 1. Total and task mean item score (MIS) – all subjects.

However, scores have been higher in some fields and on some dimensions of teaching than others. These data have contributed, along with other measures, to continuing changes in programme design. For example, the programme expanded and refined its instruction about assessment and the teaching of English language learners in response to this information, combined with other measures.

An equally important question is whether these indications of practice at the end of student teaching are related to the success of candidates in supporting student learning when they enter the profession. We turn to this question next.

Research on graduates' effectiveness

As noted earlier, the most difficult and, to many, the most important question, is how teachers' learning ultimately influences what their pupils learn. Even if teacher education students are followed into their classrooms, 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 programmes 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.

As part of the Teachers for a New Era (TNE) reform initiative at Stanford, researchers drew a sample of approximately 250 secondary teachers of mathematics, science, history/social studies, and English language arts and roughly 3500 students taught by these teachers from a set of six high schools in the San Francisco Bay Area. Because California did not at this time have a state longitudinal data system, student and teacher data had to be secured from individual schools and districts' data files, sometimes assembled by hand. The schools were from several communities and served predominantly low-income students and students of colour (for details, see Newton 2010).

Conceptualization of teacher 'effectiveness'

For the purpose of this study, only teachers of English language arts and mathematics were included. The measurement of 'value added' gains in achievement was based on the variation in pupils' test scores on the California Standards Tests (CSTs) in English language arts and mathematics, controlling for prior-year scores. Scale scores from each CST were converted to z scores based on the sample mean and standard deviation of a particular subject test. The study used ordinary least square regression analyses to predict pupils' CSTs after taking into consideration prior year's achievement (CST scores in the same subject area) and key demographic background variables (i.e. race/ethnicity, gender, free/reduced lunch status, English language learner status, and parent education). The study also controlled for school fixed effects to take into account the unobserved differences among schools that may influence teachers' measured effectiveness (e.g. school leadership, resources, parental involvement).

With these statistical controls, teacher's effectiveness was then measured by the average difference between actual scores and predicted scores for all students assigned to that teacher (i.e. the mean residual). Teachers' preparation and pathway to teaching were ascertained through surveys given to all teachers in the sample schools.

Results

Figure 2 displays the average teacher effectiveness estimates of STEP graduates versus others by years of teaching experience (i.e. >8 vs. ≤8). The average teacher effectiveness estimates for STEP graduates who had taught for more than eight years were about .30 standard deviations above the mean effectiveness of other similarly experienced teachers. The mean effectiveness estimates for STEP graduates who had taught eight or fewer years were about .14 standard deviations above those of nonSTEP graduates with similar levels of experience. Whereas STEP graduates appear to experience returns to experience beyond eight years (that is, greater effectiveness with more years of experience), the reverse was true for non-STEP teachers, for whom the less experienced cohort appeared more effective than the highly experienced group. This may be a function of improvements in the preparation of teachers generally over recent

years, which has been a goal of state policy. Figure 2. Techer effectiveness estimates – STEP vs. non-STEP.

Figure 3 displays the average teacher effectiveness estimates for alumni from different teacher education programmes and pathways. As shown in Figure 2, graduates from STEP produced higher value-added achievement gains for their students than those of the other teacher education programme groups and teachers from intern/ alternative programmes. Interestingly, these rankings for programme types track those found for the same groups of programmes on the PACT assessment instrument, which suggests that the analyses of teachers' performance at the end of their preparation programmes may be a predictor of their later effectiveness in the classroom, and programmes' capacities to help candidates learn the skills measured on PACT may enable them to support student learning more effectively. This bears further study.

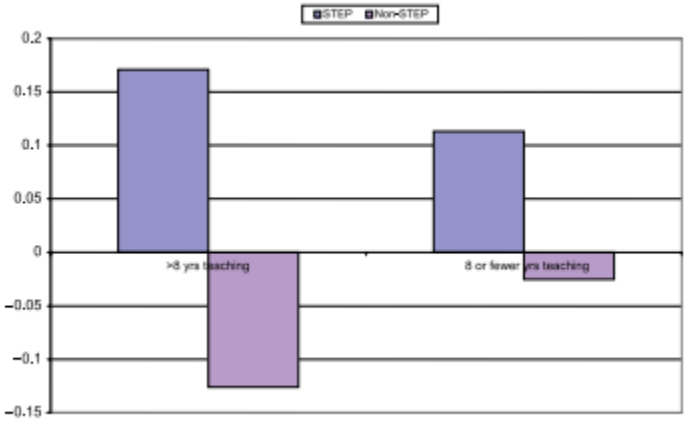

Figure 2. Teacher effectiveness estimates – STEP vs. non-STEP.

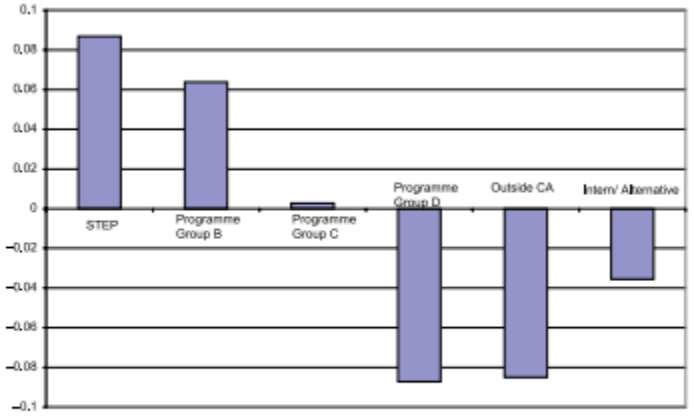

Figure 3. Estimates of high school student value-added achievement for graduates of teacher education programmes/pathways.

Conclusion

Each of the kinds of tools described here has the potential to contribute different insights to an assessment of candidates' performance and programme outcomes. Although each has

limitations, they can be powerful in the aggregate for shedding light on the development of professional performance and how various programme elements support this learning

Although there is strong press for the use of measures of teacher effectiveness as measured by student achievement gains, these are unlikely to help teacher educators improve programmes without a rich array of other tools that reveal how specific experiences support candidates in developing useful practices, and what areas of practice need more attention. Furthermore, there will be continuing concerns about the narrowness of the learning measured by standardised tests, and about the many challenges of collecting and analysing such data in ways that overcome the many technical and practical problems associated with their use (for a summary, see Braun 2005).

Thus, educators will need to develop many ways of looking at the impacts of teacher education on candidates' knowledge, skills, practices, and contributions to pupil learning. Using multiple measures and examining the relationships among them may help teacher educators develop a knowledge base for the continuous improvement of their own practice and may ultimately save the enterprise of teacher education as a whole.

Reference List

Ball, D., and D. Cohen. 1999. Developing practice, developing practitioners: Toward a practice-based theory of professional education. In Teaching as the learning profession: Handbook of policy and practice, ed. L.

Darling-Hammond and G. Sykes, 3–32. San Francisco, CA: Jossey-Bass. Berliner, D. 1986. In pursuit of the expert pedagogue. Educational Researcher 15, no. 7: 5–13.

Berliner, D. 1991. Perceptions of student behavior as a function of expertise. Journal of Classroom Interaction 26, no. 1: 1–8.

Bikle, K., and G.C. Bunch. 2002. CLAD in STEP: one programme's efforts to prepare teachers for linguistic and cultural diversity. Issues in Teacher Education 11, no. 1: 85–98.

Braun, Henry. 2005. Using student progress to evaluate teachers: A primer on value-added models. Princeton: ETS Policy Information Center.

Cochran-Smith, M. 2001. Constructing outcomes in teacher education: Policy, practice and pitfalls. Education Policy Analysis Archives 9, no. 11. http://epaa.asu.edu/v9n1.

Cole, A.L., and J.G. Knowles. 1995. University supervisors and preservice teachers: Clarifying roles and negotiating relationships. Teacher Educator 30, no. 3: 44–56.

Darling-Hammond, L. 2006a. Powerful teacher education: Lessons from exemplary programmes. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L. 2006b. Assessing teacher education: The usefulness of multiple measures for assessing programme outcomes. Journal of Teacher Education 57, no. 2: 120–38.

Darling-Hammond, L., R. Chung, and F. Frelow. 2002. Variation in teacher preparation: how well do different pathways prepare teachers to teach? Journal of Teacher Education 53, no. 4: 286–302.

Darling-Hammond, L., M. Eiler, and A. Marcus. 2002. Perceptions of preparation: Using survey data to assess teacher education outcomes. Issues in Teacher Education 11, no. 1: 65–84.

Darling-Hammond, L., A.E. Wise, and S.P. Klein. 1999. A license to teach. San Francisco, CA: Jossey-Bass.

Darling-Hammond, L., and P. Youngs. 2002. Defining 'highly qualified teachers': What does 'scientifically-based research' actually tell us? Educational Researcher 31, no. 9: 13–25.

Diamonti, Michael C. 1977. Student teacher supervision. Educational Forum 41, no. 4: 477–86.

Fetterman, D., W. Connors, K. Dunlap, G. Brower, T. Matos, and S. Paik. 1999. Stanford Teacher Education Programme 1997–98 evaluation report. Stanford, CA: Stanford University Press.

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

Haertel, E.H. 1991. New forms of teacher assessment. In Review of Research in Education, ed. Gerald Grant, 3–29. Washington, DC: American Educational Research Association.

Hammerness, K. 2006. From coherence in theory to coherence in practice. Teachers College Record 108, no. 7: 1241–65.

Hammerness, K., L. Darling-Hammond, and L. Shulman. 2002. Toward expert thinking: How curriculum case writing prompts the development of theory-based professional knowledge in student teachers. Teaching Education 13, no. 2: 221–45.

Howey, K.R., and N.L. Zimpher. 1989. Preservice teacher educators' role in programmes for beginning teachers. Elementary School Journal 89, no. 4: 450–70.

Kennedy, M. 1999. The role of preservice teacher education. In Teaching as the learning profession: Handbook of policy and practice, ed. L. Darling-Hammond and G. Sykes, 54–85. San Francisco, CA: Jossey Bass.

Kunzman, R. 2002. Preservice education for experienced teachers: What STEP teaches those who have already taught. Issues in Teacher Education 11, no. 1: 99–112.

Kunzman, R. 2003. From teacher to student: The value of teacher education for experienced teachers. Journal of Teacher Education 54, no. 3: 241–53.

McIntyre, J.D., D.M. Byrd, and S.M. Foxx. 1996. Field and laboratory experiences. In Handbook of Research on Teacher Education, ed. J.

Sikula, T.J. Buttery and E. Guyton, 171–93. New York: Macmillan. National Commission on Teaching and America's Future (NCTAF). 1996. What matters most: Teaching for America's future. New York: NCTAF.

Newton, X. 2010. Teacher effectiveness and pathways into teaching in California. Berkeley, CA: University of California at Berkeley.

Pecheone, R., and R. Chung. 2006. Evidence in teacher education: The Performance Assessment for California Teachers (PACT). Journal of Teacher Education 57, no. 1: 22–36.

Richardson, V., and D. Roosevelt. 2004. Teacher preparation and the improvement of teacher education. In Developing the Teacher Workforce, National Society for the Study of Education (NSSE) Yearbook, ed. M.A.

Smiley and D. Miretzky, 105–44. Chicago, IL: NSSE. 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, no. 2: 155–79.

Shulman, L.S. 1996. Just in case: Reflections on learning from experience. In The case for education: Contemporary approaches for using case methods, ed. J.

Colbert, P. Desberg and K. Trimble, 197–217. Boston: Allyn & Bacon. Shultz, S. 2002. Assessing growth in teacher knowledge. Issues in Teacher Education 11 no. 1: 49–64. US Department of Education. 2002. The secretary's report on teacher quality. Washington, DC: US Department of Education.

Williams, D.A., H. Ramanathan, D. Smith, J. Cruz, and L. Lipsett. 1997. Problems related to participants' roles and programmatic goals in student teaching. Mid-Western Educational Researcher 10, no. 4: 2–10.