

EVIDENCE IN TEACHER EDUCATION

THE PERFORMANCE ASSESSMENT FOR CALIFORNIA TEACHERS (PACT)

Raymond L. Pecheone
Ruth R. Chung
Stanford University

The Performance Assessment for California Teachers (PACT) was developed in response to a California State mandate (SB 2042), requiring teacher preparation programs to use performance assessments as one measure in making credentialing decisions. In this article, results are examined from statewide implementation of the PACT assessments during the first 2 pilot years. Despite the limitation of only 2 years of data, 3 years of implementation experiences have informed participating programs about how they can better support candidate learning and identify areas for examination. In addition, this research suggests that the PACT performance assessment can be used in teacher education as a valid measure of individual teacher competence for the purpose of teacher licensure and as a powerful tool for teacher learning and program improvement.

Keywords: *teacher performance assessment; teacher evaluation; teacher licensure; teacher education policy*

In 1998, California's state legislature elected to require teacher preparation programs to use standardized performance assessments in making credentialing decisions (California Commission on Teacher Credentialing, 2002). The California Commission on Teacher Credentialing contracted with the Educational Testing Service to develop a standardized performance assessment instrument to be used by all preparatory institutions but gave teacher education programs the option of using a different instrument if it met the state's standards for reliability and validity. A coalition of California colleges and universities¹ formed the Performance Assessment for California Teachers (PACT) to

develop an alternative standards-based assessment method to credential prospective teachers. Many teacher educators at these campuses were dissatisfied by the content and format of the state's teacher performance assessment, which was designed as a generic assessment that applies across all grade levels and subject areas. Specifically, the teacher performance assessment was developed as four separate and discrete performance tasks that are designed to be embedded in college or university preparation program courses. Motivated by a desire to develop an integrated, authentic, and subject-specific assessment that is consistent with the core values of member institutions while meet-

Authors' Note: Special thanks to participating institutions, the University of California Office of the President, the Flora and Sally Hewlett Family Foundation, the Hewlett Foundation, and the Morgan Family Foundation for their generous support for the PACT project work. Additional resources (handbooks and rubrics) can be found at <http://www.pacttpa.org>.

Journal of Teacher Education, Vol. 57, No. 1, January/February 2006 22-36
DOI: 10.1177/0022487105284045
© 2006 by the American Association of Colleges for Teacher Education

ing the assessment standards required by the state (California Commission on Teacher Credentialing, 2001), the PACT consortium has been working since the summer of 2002 to develop and pilot an integrated set of subject-specific assessments of teaching knowledge and skill (as defined by the California Teacher Performance Expectations).

The PACT assessments or teaching events (TEs) use multiple sources of data (teacher plans, teacher artifacts, student work samples, video clips of teaching, and personal reflections and commentaries) that are organized on four categories of teaching: planning, instruction, assessment, and reflection (PIAR). The PACT assessments build on efforts by the National Board for Professional Teaching Standards and the Interstate New Teacher Assessment and Support Consortium, which developed performance assessments for use with expert and beginning teachers. Like these earlier assessments, the focus of the PACT assessments is on candidates' application of subject-specific pedagogical knowledge that research finds to be associated with successful teaching (Bransford, Brown, & Cocking, 1999; Darling-Hammond, 1998; Fennema et al., 1996; Grossman, 1990; Porter, 1988; Shulman, 1987). What distinguishes the PACT assessments from the National Board for Professional Teaching Standards assessments is that the TE tasks are more integrated (capturing a unified learning segment), are designed to measure teacher performance at the preservice level, and have no assessment center components. Moreover, the PACT assessment system also uses a multiple measures approach to assessing teacher competence through the use of course-embedded signature assessments (described below).

Performance Assessments as Evidence in Teacher Education

Traditional measures of teachers' competency for licensing decisions have come under fire for their lack of authenticity and predictive validity (Darling-Hammond, 2001; Haertle, 1991; Mitchell, Robinson, Plake, & Knowles, 2001; Porter, Youngs, & Odden, 2001). The Na-

tional Research Council's Committee on Assessment and Teacher Quality recently noted that there is little evidence regarding the technical soundness of traditional teacher licensure tests in the published literature and little research documenting the validity of such licensure tests for identifying competent teachers or for predicting effective teaching in the classroom (Mitchell et al., 2001). In addition, tests administered by preparation programs usually vary between instructors and across institutions and, therefore, rarely provide data for large-scale analyses. In an era in which teacher education has been challenged to demonstrate its effectiveness, performance assessments have emerged not only as useful measures of teacher performance but also as a way to evaluate the quality of credential programs for state accountability systems and program accreditation.

Performance Assessments as Tools for Program Improvement and Teacher Learning

Performance assessments that include evidence from actual teaching practice have the potential to provide more direct evaluation of teaching ability. In addition, these assessments can inform programs about areas of strength and weakness in the preparation of their graduates for the purpose of program improvement and ultimately contribute to the improvement of teacher quality. There is a growing body of evidence that indicates that such assessments can better evaluate instructional practice (Mitchell et al., 2001; Porter et al., 2001) and that they can serve as powerful professional learning experiences (Anderson & DeMeulle, 1998; Darling-Hammond & Snyder, 2000; Lyons, 1998; Snyder, Lippincott, & Bower, 1998; Stone, 1998; Whitford, Ruscoe, & Fickel, 2000). In addition, several studies confirm that the assessments of the National Board for Professional Teaching Standards predict teacher effectiveness as evaluated by their students' learning gains (Bond, Smith, Baker, & Hattie, 2000; Cavalluzzo, 2004; Goldhaber & Anthony, 2004; Hattie, Clinton, Thompson, & Schmidt-Davis,

1995; Vandervoort, Amrein-Beardsley, & Berliner, 2004), although other studies do not find such conclusive evidence (Ballou, 2003; Pool, Ellett, Schiavone, & Carey-Lewis, 2001).

This article examines results from statewide implementation of the PACT during the 2002-2003 and 2003-2004 pilot years. It also describes the PACT assessment design and scoring system, summarizes the results of validity/reliability studies conducted during the first 2 pilot years, and describes ways in which the evidence drawn from the PACT has been used by teacher education programs to evaluate program effectiveness, to refine their programs, and to improve the preparation of beginning teachers. Despite the limitation of only 2 years of pilot data, there is some preliminary evidence that the implementation of the PACT assessment can be a catalyst for program change. There is also evidence that preservice teachers' learning experiences and development as teachers were enhanced by their participation in the pilot.

ASSESSMENT DESIGN

The PACT project focuses on two assessment strategies: (a) the formative development of prospective teachers through embedded signature assessments that occur throughout teacher preparation and (b) a summative assessment of teaching knowledge and skills during student teaching (the TE). In 2002, 12 colleges and universities began collaborating to identify and share exemplary curriculum-embedded assessments across programs. These embedded assessments include case studies of individual students, lesson or unit plans, analyses of student work, and observations of student teaching. The purpose of the embedded signature assessments is to provide formative feedback to the prospective teacher and teacher educators as well as to provide multiple sources of data to inform the licensure decision. Through a Web site, these assessments are being shared across the PACT institutions to both build understandings and share new approaches to teacher assessment.

The TEs are subject-specific assessments that are integrated across the four PIAR tasks. To

complete the TE, candidates must plan and teach a learning segment (i.e., an instructional unit or part of a unit), videotape and analyze their instruction, collect student work and analyze student learning, and reflect on their practice. Student teachers (all of whom are enrolled in 1-year postbaccalaureate credential programs or in 2-year internship programs) document 3 to 5 hours of their instruction of this unit (usually a week of instruction) near the end of their final student teaching placements.

The TEs are designed to measure and promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions and to stimulate teacher reflection on practice. By probing candidate thinking about student learning through the critical analysis of student work samples, the assessments provide important opportunities for mentoring and self-reflection. Additional benefits include focusing attention on the academic language² development of all students, especially English learners and native speakers of varieties of English, as well as instructional strategies that are effective with a wide range of students.

During 2002 to 2003, PACT developed assessments in five certification areas: elementary (literacy and mathematics), secondary English/language arts, secondary mathematics, secondary history/social science, and secondary science. Following each pilot year, PACT has revised the TEs in these areas in consideration of feedback received from candidates and teacher educators who had piloted the assessments. (See the appendix for an overview of the elementary literacy TE.) During the 2004-2005 pilot year, the consortium created assessments in several additional areas, including world languages, art, music, and physical education.

The PACT scoring system includes identification of subject-specific benchmark TEs and a standardized training design used to calibrate scorers and assess scorer consistency. In both pilot years, scorers were recruited by subject area and included faculty, supervisors, cooperating teachers, National Board-certified teachers and other experienced teachers. For the 1st-year pilot, a centralized scoring model was used

to get a better understanding of the training needs. Scorer training and scoring were conducted at five regional sites throughout the state. For the 2nd-year pilot, scoring was done within each local program based on a trainer-of-trainers model. Using this scoring model, local programs participated in a centralized audit process in which they drew a 20% sample of their TEs to be rescored to obtain an independent estimate of scorer agreement or disagreement in comparison to locally assigned scores.

To score the TE, scorers used a task-based scoring model. The score path follows the design of the portfolio in a sequential manner, reading and scoring each task in sequence. Each task includes multiple guiding questions (GQs) and corresponding rubrics (on a 4-point continuum) to rate each candidate's performance on each of the GQs (see Table 3 for a list of the common GQs across all subject areas). Each TE takes approximately 2 to 3 hours to score. As a result of this process, a detailed score profile is generated that provides information at the GQ level and at the PIAR task level. Individual score profiles can then be used by the candidate to develop an individual induction plan for use in California's statewide induction programs. Aggregated scores for all candidates within a program may also be used as a basis for internal and external reviews of the credential program for the purpose of program improvement, course development, and accreditation.

2003-2004 PILOT YEARS—SCORE DATA

To test the utility and viability of the PACT TE within and across subject areas, the assessment was piloted in 11 PACT programs in the 1st year and 13 PACT programs in the 2nd year. The number of piloting candidates varied across institutions, with each PACT institution purposefully selecting particular subject areas or cohorts of students to participate in each year's pilot.

Score Samples

During the 2002-2003 pilot, 395 TEs were scored at regional scoring sites. Of those 395 TEs, 163 (about 41%) were double scored to

evaluate interrater reliability. During the 2003-2004 pilot, 628 TEs were scored at local campuses. Of those TEs, 203 (about 32%) were audited (independently scored) at the centralized audit site using experienced scorers.

Score Profiles (2002-2003 Pilot Year)

This section provides an overview of scores across subject areas, summarizing how well teacher candidates did on the TE across all 11 PACT institutions that participated in the 1st-year pilot. See Table 1 for a summary of average scores across subject areas and all participating institutions. The total number of GQs for each subject area varied from 15 questions to 18 questions. However, all subject-area scoring rubrics had the same four PIAR tasks. *Subscores* refers to the scores on each of these PIAR tasks. For greater interpretability, the total scores and subscores were converted to mean item scores (MISs), which refer to the total score or subscore divided by the number of GQs in each category. Because each GQ was scored on a rubric scale of 1 to 4, each MIS falls somewhere between 1 and 4, giving a sense of how well a candidate performed across all rubric items, as well as within the PIAR categories. A score of 1 on an individual GQ means the candidate has not met the standard on that item. A score of 2 means the candidate has met the standard at the minimum level. Scores of 3 and 4 represent advanced and superior levels of performance, respectively. Scanning the MISs across the PIAR categories in Table 1, it is apparent that teacher candidates in most subject areas scored higher on the planning and instruction categories and scored lower on the assessment and reflection categories. The total MISs indicate that the average performance across subject areas met the minimum standard of performance both within the PIAR tasks and overall.

Score Profiles (2003-2004 Pilot Year)

As in the 1st pilot year, candidates scored significantly higher on the planning rubrics than on the other tasks and significantly lower on the academic language task. (Figures 1 and 2 show

TABLE 1 Mean Item Scores (MISs) by Subject Area (2002-2003 Pilot Year)

Content Area	Total MIS	Planning MIS	Instruction MIS	Assessment MIS	Reflection MIS	Academic Language MIS ^a	Total N
EL	2.34 (.597)	2.51 (.643)	2.51 (.663)	2.09 (.708)	2.16 (.843)	2.18 (.661)	123
EM	2.11 (.512)	2.24 (.544)	2.18 (.644)	1.92 (.528)	2.00 (.737)	2.00 (.592)	123
ELA	2.00 (.505)	2.15 (.573)	2.08 (.516)	1.86 (.616)	1.95 (.709)	1.89 (.524)	53
MTH	2.28 (.474)	2.54 (.493)	2.48 (.673)	1.98 (.654)	1.90 (.707)	2.05 (.629)	25
HSS	2.38 (.540)	2.44 (.593)	2.48 (.582)	2.34 (.613)	2.26 (.610)	2.37 (.615)	40
SCI	2.44 (.507)	2.52 (.529)	2.46 (.613)	2.36 (.586)	2.48 (.724)	2.61 (.557)	31

NOTE: Standard deviations in parentheses. EL = elementary literacy; EM = elementary mathematics; ELA = secondary English; MTH = secondary mathematics; HSS = secondary history–social science; SCI = secondary science.

a. The academic language MIS is the average MIS for Planning Guiding Question 5, Instruction Guiding Question 3, and Assessment Guiding Question 2, therefore, there is some overlap between the academic language MIS and other PIAR (planning, instruction, assessment, reflection) MISs.

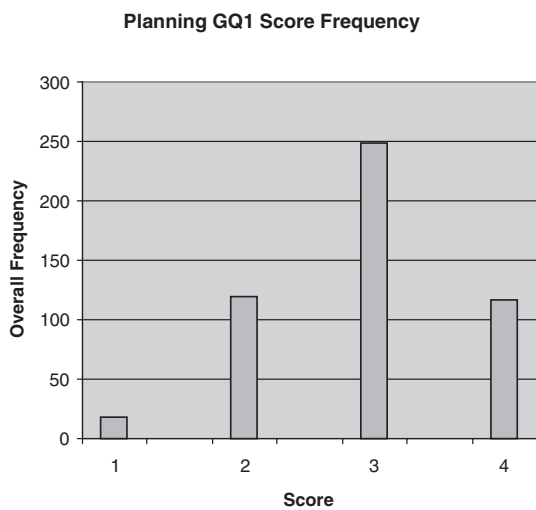


FIGURE 1: Planning Rubric 1 Scores (2004)
NOTE: GQ = guiding question.

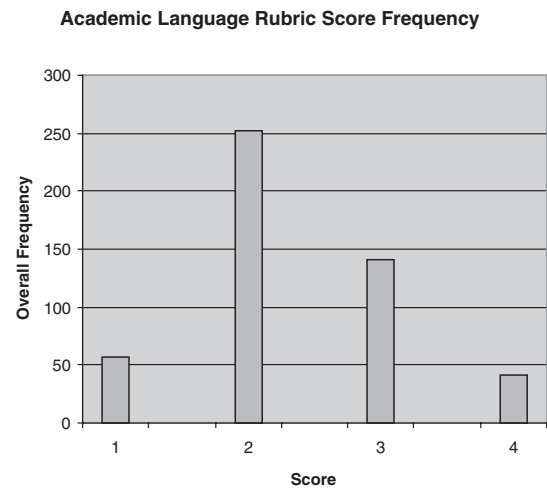


FIGURE 2: Academic Language Rubric Scores (2004)

that the distribution of candidates' scores on the Planning GQ1 and Academic Language GQ13 are dramatically different.) However, the differences in mean item subscores among the instruction, assessment, and reflection tasks are not as pronounced as they were in the previous year, and differences are not statistically significant (when scores for all subject areas are aggregated).

Table 2 (2004 pilot score data) shows higher total MISs and higher PIAR mean item subscores across subject areas than those seen in the 2003 pilot year. Although strong conclusions from only 2 years of pilot data cannot be drawn, a range of factors may speculatively account for year-to-year differences, including program improvements, smoother implementation, revisions of the TE and rubrics, or the

shift from a centralized scoring model to a local campus scoring model—which could lead to some score inflation.

Table 3 displays MISs across common sets of standardized rubrics that were used in the 2nd pilot year (2002-2003). The patterns of performance across the rubrics are similar to those of the previous year, with the planning task scores the highest and other task scores lower. Another finding consistent with the scores from the 2003 pilot is that the lowest MISs were again on the academic language item and the feedback item. Discussion among teacher educators across the consortium suggests that these lower scores may be related to both the developmental weaknesses of beginning teachers and a lack of emphasis on these skills in their program curricula (PACT standard setting meeting minutes, January 24, 2005, and February 23, 2005).

TABLE 2 Mean Item Scores (MISs) by Subject Area (2003-2004 Pilot Year)

Content Area	Total MIS	Planning MIS	Instruction MIS	Assessment MIS	Reflection MIS	Academic Language MIS	Total N
EL	2.62 (.657)	2.81 (.719)	2.53 (.827)	2.46 (.843)	2.55 (.803)	2.47 (.873)	224
EM	2.40 (.602)	2.53 (.630)	2.37 (.683)	2.28 (.790)	2.47 (.707)	2.19 (.768)	109
ELA	2.56 (.723)	2.72 (.718)	2.53 (.874)	2.44 (.852)	2.45 (.945)	2.38 (.855)	110
MTH	2.35 (.554)	2.52 (.719)	2.34 (.679)	2.30 (.672)	2.27 (.497)	1.84 (.710)	50
HSS	2.43 (.530)	2.66 (.607)	2.34 (.651)	2.31 (.682)	2.27 (.540)	2.18 (.567)	60
SCI	2.67 (.569)	2.83 (.650)	2.67 (.653)	2.53 (.684)	2.53 (.611)	2.49 (.650)	72
WL	2.64 (.270)	3.00 (.200)	2.33 (.289)	2.56 (.385)	2.50 (.500)	2.00 (.000)	3

NOTE: Standard deviations in parentheses. EL = elementary literacy; EM = elementary mathematics; ELA = secondary English; MTH = secondary mathematics; HSS = secondary history–social science; SCI = secondary science; WL = secondary world languages. Audit scores were excluded in this summary. Double scores were averaged for teaching events that were scored more than once. Calibration scores at the local campus level were each included as separate scores.

TABLE 3 Descriptives—Common Rubric Items (2003-2004 Pilot Year)

Common Rubric Items ^a —Guiding Questions	N	Mean ^b	Standard Deviation
Planning			
Access to curriculum—How does the instructional design make the curriculum accessible to the students in the class?	628	2.90	.806
Coherent instructional design—How does the instructional design reflect a coherent approach to the literacy curriculum?	627	2.76	.829
Balanced instructional design—How does the instructional design reflect a balanced approach to the literacy curriculum?	626	2.67	.809
Student needs and characteristics—How does the instructional design reflect and address student interests and needs?	627	2.63	.823
Assessment alignment—How well are the learning goals, instruction, and assessments aligned?	626	2.62	.787
Instruction			
Engagement—How does the candidate actively engage students in their own understanding of relevant skills and strategies to comprehend and/or compose text?	622	2.54	.789
Monitoring learning—How does the candidate monitor student learning and respond to student comments, questions, and needs?	617	2.50	.783
Assessment			
Whole class learning—How does the candidate's analysis of whole class learning reveal students' understanding of literacy?	625	2.49	.875
Individual learning progress—How does the candidate analyze the two students' progress over time?	625	2.56	.897
Feedback—What is the quality of oral and written feedback to the two students about literacy?	623	2.22	.957
Reflection			
Focus of reflections—To what extent did the candidate's reflections focus on student learning?	625	2.53	.855
Teaching and learning—What is the relationship between the candidate's reflections on teaching and on learning?	625	2.42	.784
Academic language—How does the candidate's planning, instruction, and assessment support academic language development?	618	2.33	.808

a. In the 2nd pilot year, all subjects had common rubrics that were tailored with subject-specific language. The guiding questions listed here are for the elementary literacy teaching event.

b. Only local scores were included in this score analysis.

Individual candidate score profiles and aggregated campus-level profiles of performance across GQs and tasks by each content area were distributed to each consortium credential program after each year of piloting and

scoring. These profiles provided information about areas of strength and weakness in teacher candidates' performance that could potentially be useful for programs in supporting individual candidate learning in areas of teaching identi-

fied as needing further development, as well as in determining ways to strengthen teacher education programs. Responses to a survey of program directors and teacher educators across the consortium campuses indicate that many participating campuses made note of candidates' weaker performances on the assessment and reflection tasks of the TE and made efforts to provide more support and guidance in completing the TE in these categories. In addition, survey responses indicate that curricular and organizational changes have been made to programs as a consequence of participating in the PACT pilot. For example, many respondents cited a greater focus on developing candidates' skills in analyzing student learning to guide future instruction, changes in course content to scaffold candidates' experiences with the TE, and an improvement in communication across program faculty, supervisors, and cooperating teachers that has led to a greater consensus about what their graduates should know and be able to do.

A comparison of candidates' responses on the Teacher Reflection Survey (described in more detail in the Learning Consequences section below) during the 2 years indicates that candidates in the 2nd-year pilot were much more likely to agree that their teacher preparation program and student teaching experiences had prepared them to complete the TE (62% agreed that their program courses had prepared them in 2003 and 84% agreed in 2004; 63% agreed that their student teaching placements had prepared them in 2003 and 90% agreed in 2004). Candidates were also more likely to rate their university professors' support higher in the 2nd pilot year (42%) than they did in the 1st pilot year (28%), whereas their ratings of support received from other credential candidates, university supervisors, and cooperating teachers remained about the same or slightly declined. These results further suggest that teacher education programs are beginning to make changes in the ways they structure and support candidates to create conditions in which the implementation of the TE can be a learning experience for both faculty and their students.

SUMMARY OF VALIDITY AND RELIABILITY ANALYSES FOR THE TE

The utility of the scores presented above for licensure and program improvement depends on the degree to which the TE is valid and reliable. Therefore, evaluating the validity of the instrument for its ability to accurately and fairly measure the teaching skills of prospective teachers has been a critical activity of the PACT consortium. Validity in this instance refers to the appropriateness, meaningfulness, and usefulness of evidence that is used to support the decisions involved in granting an initial license to prospective teachers. As previously noted, the PACT assessment system was designed to meet a legislative requirement that prospective teachers demonstrate proficiency on a teaching performance assessment to be eligible for a preliminary credential. With this purpose in mind, validity studies were conducted to examine the extent to which candidate performance on the TE can accurately differentiate between effective candidates (those who meet or exceed California's Teaching Performance Expectations [TPEs]) and ineffective candidates (those who do not meet the TPEs at a minimum level). This section highlights results from the studies that have been conducted to date to determine the validity and reliability of the TE.

Content Validity

Content validity addresses the question, How well does the content of the PACT assessments represent a particular domain of professional knowledge or skills? Historically, content validity coupled with a job analysis has been the primary source of validity used to support teacher licensure assessment (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985, 1999; Wilkerson & Lange, 2003). In the process of establishing content validity, the test developer frames the assessment on the knowledge and skills that are represented in the teaching standards. The California TPEs, established by policy makers, teachers, teacher educators, and administra-

tors, are based on a statewide job analysis that forms the basis for credentialing prospective teachers. Therefore, the structure, substance, and representativeness of the TE tasks were examined in relationship to the TPE domains. One study examines the alignment of the TE tasks to the California TPEs by breaking down the TE task prompts and examining their alignment with different elements of the TPEs (Jones & Ford-Johnson, 2004). In addition, an expert committee of teacher educators who participated in the development and design of the assessments was also asked to judge the extent to which the content of the TEs was an authentic representation of important dimensions of teaching (according to their professional judgment). Overall, the findings across both content validity activities suggest a strong linkage between the TPE standards, the TE tasks, and the skills and abilities that are needed for safe and competent professional practice.

Bias and Fairness Review

The initial bias review followed guidelines put forth by the Educational Testing Service (2002) for conducting bias/sensitivity reviews of assessments. The bias/sensitivity review process focused specifically on the TE handbooks and rubrics used in each certification area to evaluate the text for offensive or potentially offensive language and to identify any areas of bias relating to race, gender, ethnicity, or cultural-linguistic background. The process entailed training teacher educators who have a background in examining test bias to independently evaluate the PACT assessment materials and collectively determine whether sources of bias were evident. Panel participants used a structured reporting form to record their responses to the PACT materials. Findings from this process are used to flag areas of potential bias that are used to inform subsequent revisions of the TE handbooks, rubrics, and scoring process.

Second, the performance of candidates on the PACT assessment was examined to determine if candidates performed differentially with respect to specific demographic characteristics.

To test for fairness across these demographic indicators, an ANOVA or *t*-test methodology was used. For the 2002-2003 pilot year, no significant differences were found in total MISs between candidates with different demographic characteristics, including candidates with different percentages of English language learners in their teaching placement and candidates of different ethnicities.³ The difference in scores between candidates whose primary language is English versus another language was marginally significant (.065), with candidates whose primary language is English scoring .20 points higher on average in the total MISs. However, because standard setting was not conducted in the 1st year, we do not know to what extent these mean score differences might have a differential impact (in terms of pass rates) on the two groups.

For the 2003-2004 pilots, there were no significant differences in scores by race/ethnicity of candidates, percentage of English language learner students in candidates' classrooms, and the socioeconomic status of the classroom. There were, however, significant differences between male and female candidates, with females scoring higher, and between candidates teaching in schools in different socioeconomic contexts, with candidates in suburban schools scoring higher than those in urban or inner-city schools. One hypothesis for the latter finding is that candidates teaching in urban settings were more likely to report constraints on their teaching decisions related to district-mandated curricula. Analysis of scores indicates that higher levels of reported constraints were associated with lower scores on the TE. In addition, the instructional contexts in urban areas are often more challenging and generally require greater command of teaching skills to meet students' diverse learning needs. As a result of these findings, PACT will continue to monitor and reexamine the scorer training process as well as the design features of the TE such as GQs, rubrics, and benchmarks. If sources of bias are identified because of socioeconomic status or other variables, then modifications will be made to the assessment system to address the problem areas.

Construct Validity

Another aspect of validity is the examination of construct validity (Cronbach, 1988). Construct validity focuses on examining the meaning of PACT scores in terms of psychological or pedagogical constructs that underlie the assessment. Constructs in this context permit categorization and description of some directly observable behavior (Crocker & Algina, 1986). Factor analyses were conducted on the score data from both pilot years. In the 2002-2003 pilot year, three factors emerged from the elementary literacy score data with assessment/reflection, instruction, and planning items composing each factor. Similar results emerged from the common rubric item scores across all subject areas. In the 2003-2004 pilot year, two factors emerged from the elementary literacy and the common rubric item scores, with the first factor composed of planning and instruction items and the second factor composed of assessment and reflection items. These results suggest that the assessment tasks (PIAR) are generally supported by the factor analyses and appear to be meaningful categories that represent significant domains of teaching skill.

Score Consistency and Reliability

In these analyses, we examined the consistency across pairs of scores by computing interrater agreement percentages within each subject area. During the 1st pilot year, we examined the level of agreement between scores of double-scored TEs. During the 2nd pilot year, we examined the level of agreement between campus-assigned scores and audit scores for the same TEs. In the 1st pilot year, we found that 90% of score pairs were exact matches or within 1 point. In the 2nd pilot year, we found that 91% of scores were exact matches or within 1 point. In the 3rd pilot year, score reliability will be further examined by conducting a generalizability study to break down sources of error by raters, tasks, and occasions, as well as to determine decision consistency with regard to the cut score.

Concurrent Validity

Studies of concurrent and/or criterion-related validity are seldom included in the validation of licensure tests (Poggio, Glasnapp, Miller, Tollefson, & Burry, 1986). One of the major complications of these studies is the need to find adequate criterion measures that can be used to measure candidate effectiveness on the same or a similar domain of teaching skills and abilities. Two studies have been conducted to explore the relationship of the PACT scores to other indicators of candidates' competence: (a) a comparison of scorers' analytic ratings with their holistic ratings of the candidate and (b) the degree to which candidates' supervisors or instructors agreed or disagreed with the PACT ratings.

Study 1: Comparing Analytic and Holistic Scorer Ratings

The research hypothesis for this study was that a candidate's MIS on the TE should be consistent with the raters' recommendation for a credential based on a holistic rating of the candidate's performance. In each pilot year, scorers were asked to step back after scoring a TE and holistically evaluate the candidate's performance based on the following question and rating scale:⁴

We would like to collect your impression of the performance in the Teaching Event independent of the rubric ratings. If the evidence of teaching practice in this Teaching Event was typical of a candidate's current level of practice, what would be your recommendation with respect to awarding them a teaching credential?

- 1 = "Would not recommend for a Teaching Credential (candidate's areas of weakness cause concerns for being the teacher of record)"
- 2 = "Recommendation for a Teaching Credential (has areas of strength that will carry candidate while s/he works on areas that need improvement)"
- 3 = "Strong recommendation for a Teaching Credential (solid foundation of beginning teaching skills)"
- 4 = "Strong recommendation for a Teaching Credential (exceptional performance for a beginner)"

When these holistic scores are compared to the analytic scores raters had given the same

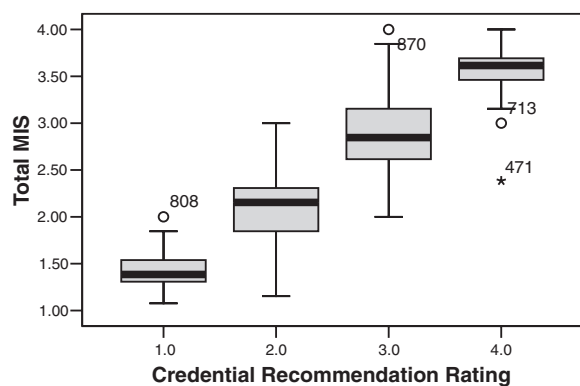


FIGURE 3: Box Plot—Total Mean Item Scores (MISs) by Credential Recommendation Rating (All Subject Areas)—2004 Pilot Year

candidates, there is strong agreement between the analytic scoring of the TE and a holistic judgment about whether a candidate should be granted a credential. These two sets of ratings are certainly not independent because they are drawn from the same body of evidence by the same rater. However, comparison of analytic and holistic scores is useful because it lends some support to the assertion that the domains of teaching skill captured in the rubrics are valued by raters as components of minimally competent teaching practice that would be needed to qualify for a credential. An examination of the MISs for TEs of candidates falling in each of the four credential recommendation holistic ratings indicates that there were large differences (significant at the .001 level) in the average total MISs and average PIAR mean item subscores across the holistic ratings. (See Figure 3 for a box plot that shows the distribution of total MISs for each holistic rating.)

Study 2: Criterion Validity

This study, conducted after the 1st pilot year, asked supervisors and faculty members familiar with candidates to what extent (on a 5-point Likert-type scale) they agreed with the scores candidates received on the subject-specific rubrics. A total of 152 score evaluations were completed and returned from supervisors and faculty at seven different consortium programs

and across all subject areas (note that these score report evaluations had no bearing on candidates' success in earning a credential because they were solicited after candidates had graduated from the program). Most of the faculty and supervisors completing these evaluations had not participated directly in the PACT pilot and had little at stake in establishing the validity of the TE. In fact, many teacher educators were disinclined to support the new state teacher performance assessment requirement. Thus, there appears to be little or no reason for these respondents to have had a bias in favor of agreeing with the TE scores. Overall, 90% of the faculty or supervisors agreed with most of the TE scorers' candidate ratings on the 15 to 17 GQs that compose the TE rubrics, and 72% agreed with a large majority of the ratings on the GQs. The findings from this confirmatory analysis suggest that the TE is a valid representation of candidates' teaching skills as evaluated by faculty/supervisors who were most familiar with the students' teaching skills. Thus, in Studies 1 and 2 described above, the teaching knowledge and skills captured by the TE appears to be a credible and valid measure of a teacher's competence in relationship to similar judgments of competency by faculty/supervisors that were most familiar with the candidates' teaching.

Learning Consequences

A driving principle in the design and development of the PACT assessment was that the act of putting together a TE would significantly influence the way candidates think about and reflect on their teaching because of the design and formative nature of the assessment. This hypothesis is supported by mounting evidence that teachers who participate in the National Board for Professional Teaching Standards assessments find developing a portfolio is one of the most beneficial and career-altering activities that they ever engaged in.⁵ Several studies have been undertaken to examine the learning consequences of the PACT TE, including (a) the impact of the TE on preservice teacher learning (Chung, 2004, 2005) and (b) the effects of local implementation decisions on candidates'

experiences with the TE (Pecheone, Chung, Whittaker, & Sloan, 2005).

In the 1st-year pilot, 527 piloting candidates completed and submitted a short survey (Candidate Reflection Survey) that was included as an appendix in the TE handbook. This survey queried PACT piloting candidates about perceptions of their preparation to complete the TE, the support they had received in completing the TE, and whether important aspects of their teaching were assessed by the TE. Candidates were more likely to report that the TE assessed important elements of their teaching knowledge and skill when they felt better supported and better prepared for completing the TE. Of those who felt well supported, nearly 70% believed the TE assessed important aspects of their teaching knowledge and skill; of those who felt well prepared by course work and student teaching placements, approximately 85% believed the TE assessed important aspects of their teaching knowledge and skill.

In the 2nd-year pilot, a more extensive survey asked candidates about the learning value of various aspects of the TE. Of 590 respondents, 60% agreed or strongly agreed with the statement, "I learned important skills through the process of completing the Teaching Event." In addition, 72% agreed that the PACT had improved their ability to reflect on their teaching, and 64% agreed that the PACT had improved their assessment of student learning. Again, we found that candidates' reports of learning from constructing the TE were positively associated (significant at the .001 level) with reported levels of support and preparation provided by program courses and student teaching placements.

Finally, there seems to be an association between candidates' learning experiences and their scores on the TE. (Candidates had no knowledge of the scores they received on the TE because the survey was submitted with the TE.) Candidates who strongly agreed that they had learned important skills from the TE scored significantly higher than those who strongly disagreed that they had learned important skills. In addition, candidates who had longer student teaching placements and candidates who strongly agreed that their course work prepared

them for the TE scored significantly higher on the assessment.

These survey results (along with case studies of candidates engaged in the process of completing the TE; see Chung, 2004, 2005) provide support for the beneficial learning consequences of the TE, suggesting that students who receive targeted support in their development of the TE view their experience more positively and report that the process of constructing their TEs strengthened their teaching. These results also have important implications for implementation decisions made at the campus level. The more support provided to candidates by program faculty, supervisors, and master teachers, the more likely candidates are to have positive learning experiences and by extension, stronger performances on the TE.

That candidates' reported learning experiences and their actual scores on the TE varied by program and by the conditions in which they completed the TE suggests that a performance assessment like the PACT can serve as one indicator of beginning teachers' preparation to teach as well as an indicator of a candidate's opportunity to learn. The findings of this research also provide important feedback to credential programs about ways that they can better support prospective teachers' learning experiences and strengthen their preparation in specific areas of teaching knowledge or skills.

CONCLUSION

During the past decade, there has been some promising work in the development of performance assessment systems that are standards based and can be used as a component of teacher credentialing or as a component of program submissions for national accreditation (National Council for Accreditation of Teacher Education or Teacher Education Accreditation Council). In addition to California's teacher performance assessment system, at least one other state, Connecticut, has successfully implemented a teacher performance assessment similar to the PACT TE for the purpose of initial teacher licensure and to support teacher induction (Connecticut Department of Education, 2003). The implementation of the PACT TE in a

number of campuses in California permits the examination of program differences in candidate performance within and across credential programs and provides opportunities for collaboration across universities to both address areas of candidate weakness and to share best practices. In addition, participation in the PACT project related to the design, scoring, and implementation of the assessment has led many programs to begin professional dialogues within and across program faculty/supervisors about what constitutes effective teaching at the preservice level and about what graduates should know and be able to do on completion of a credential program. These dialogues have led programs to reexamine the way they support and prepare candidates to teach. Beyond accountability purposes, a well-designed performance assessment system has the potential to be a powerful tool in providing meaningful feedback to individual candidates and to evaluate the impact of teacher education programs.

However, use of the PACT assessment to inform state licensure and/or program accreditation will first require that the reliability and validity of the TE for assessing beginning teacher performance be established. In addition to the studies summarized in this article, a standard setting study is currently under way to determine how simulated cut scores would affect pass rates. However, we have yet to see the policy impact of PACT implementation on teacher credentialing because it has not yet been fully implemented in California as a high-stakes assessment for teacher licensure. In anticipation of state approval, the PACT assessment system has been structured to include multiple measures (i.e., the TE in combination with embedded signature assessments) that may be used to support decisions on teacher credentialing and/or program completion. Furthermore, a longitudinal study of the predictive validity of the PACT assessment is currently being planned for implementation in 2005 to 2006. In this study, credentialed teachers who have completed the PACT will be followed into their 1st years of teaching to examine their teaching practices and the relationship of their scores on

the PACT assessment with their students' achievement gains. If it can be shown that performance on the TE significantly correlates with student learning (predictive validity), then PACT would have a compelling evidence base to support the use of the TE to credential prospective teachers and/or as a component of program accreditation. Up to now, the validity of licensure tests almost exclusively has been based on studies of content and job-related validity because the collection of evidence of predictive validity has been viewed as too difficult and unreliable.

Another challenge in implementing a teacher performance assessment for licensure is that the costs associated with developing the PACT assessment system have been significant. Many of the PACT consortium participants have contributed both financially (through in-kind donations and reimbursement of travel and development costs) and through a reallocation of faculty time. The cost of developing the PACT assessments could not have been accomplished without member contributions and financial support from the University of California Office of the President and private foundations (in the absence of state financial support). In addition, once the assessments have been approved by the state for use, individual campuses will still need to dedicate resources to implement the assessment system. The extent of these costs will depend on the implementation choices of individual campuses in terms of reallocating faculty resources and time, integrating the assessment into existing teacher education program components, and incorporating the scoring of the PACT TE into the contracted responsibilities of supervisors and cooperating teachers.

Despite legitimate concerns about the costs of implementing performance-based assessments, the alternative is to judge teacher competence solely on the basis of standardized multiple-choice tests of content and/or pedagogical knowledge. State licensing policies that ignore the assessment of teaching performance will, in effect, serve to trivialize and undermine our understanding of the complexity of teachers' work and diminish the critical role of teacher education in preparing teachers. In this current

policy environment, it appears that it will not be long before states begin to develop alternative teacher credentialing routes that depend solely on test scores and bypass any clinical demonstration of teaching skill. In contrast, a well conceptualized teacher assessment system that

incorporates multiple sources of data, including an assessment of teaching performance, has the potential to provide the evidence needed to demonstrate the significant contribution of teacher education on teaching performance and ultimately on student learning.

APPENDIX

Overview of Teaching Event Tasks for Elementary Literacy (2004-2005 Pilot Year)

<i>Teaching Event Task</i>	<i>What To Do</i>	<i>What To Submit</i>
1. Context for learning (TPEs 7, 8)	<ul style="list-style-type: none"> √ Provide relevant information about your instructional context and your students as learners of literacy within the learning segment. 	<ul style="list-style-type: none"> <input type="checkbox"/> Context form <input type="checkbox"/> Context commentary
2. Planning instruction & assessment (TPEs 1, 2, 3, 4, 6, 7, 9, 10)	<ul style="list-style-type: none"> √ Select a learning segment of three to five lessons that develops students' ability to comprehend and/or compose text and that develops their reading, writing, and use of academic language. √ Create an instruction and assessment plan for the learning segment. √ Write a commentary that explains your thinking behind the plans. √ Record daily reflections to submit in the reflection section of the teaching event. 	<ul style="list-style-type: none"> <input type="checkbox"/> Overview of plans for learning segment form <input type="checkbox"/> Instructional materials <input type="checkbox"/> Planning commentary
3. Instructing students & supporting learning (TPEs 1, 2, 4, 5, 6, 7, 10, 11)	<ul style="list-style-type: none"> √ Review your plans and prepare to videotape your class. Identify opportunities for students to use relevant skills and strategies to comprehend and/or compose text. √ Videotape the lesson you have identified. √ Review the videotape to identify one or two video clips portraying the required features of your teaching. The total running time should not exceed 15 minutes. √ Provide a copy of the plan for the lesson from which the clip(s) were taken. √ Write a commentary that analyzes your teaching and your students' learning in the video clip(s). 	<ul style="list-style-type: none"> <input type="checkbox"/> Video clip(s) <input type="checkbox"/> Video label form <input type="checkbox"/> Lesson plan <input type="checkbox"/> Instruction commentary
4. Assessing student learning (TPEs 2, 3)	<ul style="list-style-type: none"> √ Select one student assessment from the learning segment and analyze student work using evaluative criteria (or a rubric). √ Identify three student work samples that illustrate class trends in what students did and did not understand. √ Write a commentary that analyzes the extent to which the class met the standards/objectives, analyzes the individual learning of two students represented in the work samples, and identifies next steps in instruction. 	<ul style="list-style-type: none"> <input type="checkbox"/> Student work samples <input type="checkbox"/> Evaluative criteria or rubric <input type="checkbox"/> Assessment commentary
5. Reflecting on teaching & learning (TPEs 12, 13)	<ul style="list-style-type: none"> √ Provide your daily reflections. √ Write a commentary about what you learned from teaching this learning segment. 	<ul style="list-style-type: none"> <input type="checkbox"/> Daily reflections <input type="checkbox"/> Reflective commentary

NOTE: TPE = teaching performance expectation. Complete handbooks and rubrics can be found at <http://www.pactpa.org>.

NOTES

1. The Performance Assessment for California Teachers consortium was initially composed of the following 12 universities: University of California–Berkeley, University of California–Los Angeles, University of California–San Diego, University of California–Santa Cruz, University of California–Santa Barbara, University of California–Riverside, University of California–Davis, University of California–Irvine, San Jose State University, San Diego State University, Stanford University, and Mills College. During the 2003–2004 academic year, 4 additional institutions joined the consortium: San Francisco State University, Sacramento State University, the San Diego City Schools Intern Program, and the University of Southern California. California State University–Dominguez Hills joined in 2005.

2. Components of academic language, from more general to more specific, include such things as formal oral presentations, writing genres, comprehension or construction of texts, subject-specific vocabulary, language functions associated with tests for specific purposes within the particular academic subject, and organizational signals of different text structures. Support for developing academic language might include one or more of the following: modeling of strategies for comprehending or constructing texts, explicit communication of the expected features of oral or written texts (e.g., using rubrics, models, and frames), use of strategies that provide visual representations of content while promoting literacy development (e.g., graphic organizers), vocabulary development techniques (context cues, categorization, analysis of word parts, etc.), opportunities to work together with students with different kinds of language and literacy skills, and so forth.

3. No significant differences were found between scores received by White, Asian, and Hispanic teacher candidates. Other ethnicities had low sample sizes and, thus, score data for these groups could not be validly analyzed.

4. Similar studies were conducted in both years of implementation and yielded similar results. The results from the 2004 pilot year data are presented here.

5. Relevant studies can be found at <http://www.nbpts.org/research>.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.
- Anderson, R. S., & DeMeulle, L. (1998). Portfolio use in twenty-four teacher education programs. *Teacher Education Quarterly*, 25(1), 23–32.
- Ballou, D. (2003). Certifying accomplished teachers: A critical look at the National Board for Professional Teaching Standards. *Peabody Journal of Education*, 78(4), 201–219.
- Bond, L., Smith, T., Baker, W. K., & Hattie, J. A. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro: University of North Carolina–Greensboro, Center for Educational Research and Evaluation.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school* (Committee on Developments in the Science of Learning, Commission on Behavioral and Social Sciences and Education, National Research Council). Washington, DC: National Academy Press.
- California Commission on Teacher Credentialing. (2001). *Assessment quality standards* (Updated July 21, 2004). Retrieved from <http://www.ctc.ca.gov/educator-standards/AssessmentQualityStds-CAT-E.doc>
- California Commission on Teacher Credentialing. (2002). *SB2042: Professional preparation programs—Teaching performance assessment* (Updated February 1, 2002). Retrieved from http://www.ctc.ca.gov/SB2042/TPA_FAQ.html
- Cavalluzzo, L. C. (2004, November). *Is National Board certification an effective signal of teacher quality?* Paper presented at the Consortium for Policy Research in Education Conference on teacher compensation and evaluation, Chicago. Available from <http://www.cna.org>
- Chung, R. R. (2004, April). *The Performance Assessment for California Teachers (PACT) and preservice teachers: Under what conditions do student teachers learn from a performance assessment?* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Chung, R. R. (2005). *The Performance Assessment for California Teachers (PACT) and beginning teacher development: Can a performance assessment promote expert teaching practice?* Unpublished dissertation study, Stanford University, CA.
- Connecticut Department of Education. (2003). *Technical report of the B.E.S.T. program* (Division of Evaluation and Research Internal Document).
- Crocker, L., & Algina, J. (1986). *Introduction to a classical and modern test theory*. Fort Worth, TX: Holt, Rinehart & Winston.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.
- Darling-Hammond, L. (1998). Teachers and teaching: Testing policy hypotheses from a National Commission report. *Educational Researcher*, 27(1), 5–15.
- Darling-Hammond, L. (2001). Standard setting in teaching: Changes in licensing, certification, and assessment. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 751–776). Washington, DC: American Educational Research Association.
- Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, 16(5–6), 523–545.
- Educational Testing Service. (2002). *Standards for quality and fairness*. Princeton, NJ: Author.

- Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education*, 27(4), 403-434.
- Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Seattle: University of Washington.
- Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York: Teachers College Press.
- Haertle, 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.
- Hattie, J., Clinton, J., Thompson, M., & Schmidt-Davis, H. (1995). *Identifying highly accomplished teachers: A validation study* (National Board for Professional Teaching Standards Technical Advisory Group Research Report). Greensboro: University of North Carolina-Greensboro, Center for Educational Research and Evaluation.
- Jones, P., & Ford-Johnson, A. (2004, April). *An examination of categorical versus interpretive scoring rubrics*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.
- Lyons, N. P. (1998). Reflection in teaching: Can it be developmental? A portfolio perspective. *Teacher Education Quarterly*, 25(1), 115-127.
- Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, DC: National Academy Press.
- Pecheone, R., Chung, R. R., Whittaker, A., & Sloan, T. (2005, February). *Performance assessment in preservice teacher education: Lessons from the implementation of the Performance Assessment for California Teachers (PACT)*. Paper presented at the annual meeting of the American Association for Colleges of Teacher Education, Washington, DC.
- Poggio, J. P., Glasnapp, D. R., Miller, M. D., Tollefson, N., & Burry, J. A. (1986). Strategies for validating teacher certification tests. *Educational Measurement Issues and Practice*, 5(2), 18-25.
- Pool, J. E., Ellett, C. D., Schiavone, S., & Carey-Lewis, C. (2001). How valid are the National Board of Professional Teaching Standards assessments for predicting the quality of actual classroom teaching and learning? Results of six mini case studies. *Journal of Personnel Evaluation in Education*, 15(1), 31-48.
- Porter, A., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and uses. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259-297). Washington, DC: American Educational Research Association.
- Porter, A. C. (1988). *Understanding teaching: A model for assessment*. East Lansing: Michigan State University, Institute for Research on Teaching.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.
- Snyder, J., Lippincott, A., & Bower, D. (1998). The inherent tensions in the multiple uses of portfolios in teacher education. *Teacher Education Quarterly*, 25(1), 45-60.
- Stone, B. A. (1998). Problems, pitfalls, and benefits of portfolios. *Teacher Education Quarterly*, 25(1), 105-114.
- Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004, September 8). National Board certified teachers and their students' achievement. *Education Policy Analysis Archives*, 12(46). Retrieved November 19, 2004, from <http://epaa.asu.edu/epaa/v12n46/>
- Whitford, B. L., Ruscoe, G., & Fickel, L. (2000). Knitting it all together: Collaborative teacher education in Southern Maine. In L. Darling-Hammond (Ed.), *Studies of excellence in teacher education: Preparation at the graduate level* (pp. 173-257). New York/Washington, DC: National Commission on Teaching and America's Future, American Association of Colleges for Teacher Education.
- Wilkerson, J. R., & Lange, W. S. (2003). Portfolio, the Pied Piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives*, 11(45). Retrieved July 30, 2005, from <http://www.asu.edu/epaa/v11n45/>

Raymond L. Pecheone is Executive Director of the PACT project and the School of Redesign Network at Stanford University. He specializes in teacher performance assessment, teacher licensing, school restructuring, and authentic assessment.

Ruth R. Chung is a postdoctoral scholar at Stanford University. She specializes in teacher learning, preservice teacher education, teacher performance assessment, and classroom assessment.