

TECHNICAL REPORT

of the

**Performance Assessment for California Teachers (PACT):
Summary of Validity and Reliability Studies
for the 2003-04 Pilot Year**

March 2007

Prepared by
Raymond L. Pecheone, Ph.D.
Ruth R. Chung, Ph.D.

© PACT Consortium, 2007

**Performance Assessment for California Teachers (PACT):
Summary of Validity and Reliability Studies for the 2003-04 Pilot Year
March 2007**

CONTENTS

Executive Summary	4
Introduction & Assessment Design	8
Scoring System	12
Proposed Scoring System Design.....	16
2003-04 Pilot Year – Score Summary	20
Summary of Validity and Reliability Analyses for the Teaching Event.....	24
Content Validity.....	25
Bias and Fairness Reviews.....	27
Construct Validity.....	30
Criterion-Related Concurrent Validity.....	32
Consistency and Reliability.....	35
Assessor Reliability.....	38
PACT Standard Setting 2005-2007	39
Summary.....	48
Bibliography/References	49
Appendices.....	51
Appendix A. Overview of Teaching Event Tasks (Elementary Literacy) & Sample Rubrics for Elementary Literacy (2003-04 Pilot Year).....	51
Appendix B. Relationship of TPEs to Guiding Question.....	58
Rubrics (2003-04 Pilot)	
Appendix C. PACT Bias and Sensitivity Review – Fall 2003.....	59
Appendix D. Timeline of PACT Activities & Participants.....	66
Appendix E. Jones, P. (2005). An examination of Teacher Preparation Program Standard 19a Requirements and the Performance Assessment for California Teachers	
Appendix F. Handbook for Implementing PACT Scorer Training	
Appendix G. Statement of Intended Uses of the PACT Teaching Event Scores	
Appendix H. Elementary Literacy Teaching Event Candidate Handbook & Elementary Literacy Teaching Event Rubrics (2006-07 pilot)	

TABLES

Table 1	Total Teaching Events Scored by Subject Area (2003-04 Pilot Year)	21
Table 2	Demographic Characteristics – PACT Candidate Survey (2003-04)	22
Table 3	Mean Item Scores by Subject Area (2003-04 Pilot Year)	23
Table 4	Descriptives –Rubric Scores (2003-04 Pilot Year)	24
Table 5	TPE Elements Represented in the Teaching Event	26
Table 6	EL - Rotated Component Matrix (2003-04 Pilot Year)	31
Table 7	Correlations between Teaching Event Task Mean Item Scores (PIAR) - 2003-04 Pilot Year	32
Table 8	Holistic Rating by Subject Area (2003-04 Pilot Year)	34
Table 9	Pass/Fail Decision by Subject Area based on Final Adopted Passing Standard (adapted to account for differences in rubric structures) – 2003-04 Pilot Year	34
Table 10	Crosstabs – Pass/Fail Decision and Holistic Rating	35
Table 11	Campus Scores – Audit Scores Consensus Estimate (2003-04 Pilot Year)	37
Table 12	Assessor Reliability and SES of the Teaching Event Tasks 2003-04 Pilot Year	39

Executive Summary

Under California's new licensing system (enacted by SB 2042 in 1998), teacher candidates in teacher preparation programs are required to pass a teaching performance assessment (TPA) to earn a preliminary license. Professional teacher preparation programs may use the California Commission on Teacher Credentialing (CCTC) California TPA (developed by the CCTC and ETS) or may develop an alternative assessment that meets the CCTC's Assessment Quality Standards (CCTC, 2002). Although the implementation of the requirement was delayed for fiscal reasons, both the California TPA and the alternatives continue to be piloted. The Performance Assessment for California Teachers (PACT) consortium, comprised of pre-service teacher preparation programs throughout the state, has designed subject-specific portfolio assessments that have been piloted since the 2002-03 academic year. (The PACT consortium was initially composed of the following 12 universities: UC Berkeley, UCLA, UC San Diego, UC Santa Cruz, UC Santa Barbara, UC Riverside, UC Davis, UC 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, California State University-Sacramento, the San Diego City Schools Intern Program, and the University of Southern California. In the summer of 2005, an additional campus, CSU-Dominguez Hills, joined the consortium; in June 2006, the University of the Pacific also joined; and in October 2006, CSU-Northridge joined the PACT consortium.) As of February 2007, the PACT consortium included 27 universities. New members include: Sonoma State University, CSU-Channel Islands, CSU-Monterey Bay, University of San Diego, Pepperdine University, Notre Dame de Namur University, Saint Mary's College of California, and Holy Names University. The proportion of newly licensed teachers in the state credentialed by PACT consortium institutions is 29.2% (8,017 teachers out of 27,467 newly credentialed teachers in California, based on the 2004-05 Title II report).

This report describes the design of the PACT assessments ("Teaching Events") and summarizes the results of the second year of piloting the PACT Teaching Events (TEs), and presents relevant assessment validity and reliability studies that have been conducted. The validity and reliability studies described in this report include those that examine the content validity of the Teaching Event, bias and fairness of the assessment based on scores for different groups, construct validity based on factor analyses, concurrent validity focusing on decision consistency, and score reliability and consistency.

Assessment Design. The PACT Teaching Event is an "evidence-based system" that uses multiple sources of data - teacher plans, teacher artifacts, student work samples, video clips of teaching and personal reflections and commentaries. The Teaching Events are subject-specific assessments that are integrated across four tasks: planning, instruction, assessment, and reflection (PIAR) (with a focus on Academic Language¹ embedded

¹ Academic Language is defined as "the language needed by students to understand and communicate in the academic disciplines." Academic language includes such things as specialized vocabulary, conventional text structures within a field (e.g., essays, lab reports) and other language-related activities typical of

across the tasks). For each Teaching Event, candidates must plan and teach a learning segment of 3-5 hours of instruction (i.e., an instructional unit or part of a unit), videotape and analyze their instruction, analyze student learning, and reflect on their practice. The Teaching Events are designed to measure and promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions and reflecting on practice.

Proposed Scoring System. The scoring system will include both a local and centralized scoring model. In most years, scoring will be conducted at each local campus by a group of subject-specific trainers who will be trained centrally each year. They will train, calibrate, and monitor scorers and oversee the local scoring process, including implementation of a plan for double scoring selected Teaching Events. All failing and borderline Teaching Events will be double scored and checked by a chief trainer to confirm the decision. An additional random 10 percent sample stratified across scores levels 2-4 will be double scored by local scorers. The consistency of local scoring will be managed through a centralized audit of 10 percent of local scores, with intervention with campuses that are identified as producing unreliable scores. Every third year, a central standardized scoring model will be used to provide another check on the consistency of training and the scoring process and the reliability and validity of scores.

Sample. During the 2003-04 pilot year, about 700 candidates submitted Teaching Events, and 625 of those were scored at local campuses by raters who had been trained by PACT trainers. Additionally, 146 of those Teaching Events (about 23%) were audited at a centralized Score Audit session.

Score Profiles. The Teaching Events are scored along five major tasks or domains: Planning, Instruction, Assessment, Reflection, and Academic Language. Scores from the 2003-04 pilot indicate that candidates across all subject areas tended to perform at a higher level on the Planning and Instruction tasks than on the Assessment and Reflection tasks. In addition, candidates tended to perform at a lower level of performance on the Academic Language related rubrics.

Content Validity. Teacher educators who participated in the development and design of the assessments were asked to judge the extent to which the content of the Teaching Events was an authentic representation of important dimensions of teaching. Another study examined the alignment of the TE tasks to the California Teaching Performance Expectations (TPEs). Overall, the findings across all 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.

classrooms, (e.g., expressing disagreement, discussing an issue, asking for clarification). Academic language includes both productive and receptive modalities. For additional examples of academic language across the content areas, see the Academic Language rubrics on each of the scoring rubrics. See also pp.43-45 of **Appendix F. Handbook for Implementing PACT Scorer Training** for additional explanations of Academic Language.

Bias and Fairness Reviews. The initial bias review followed guidelines put forth by Educational Testing Service (2002) for conducting bias/sensitivity reviews of assessments. The bias/sensitivity review process focused specifically on the Teaching Event 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 due to race, gender, ethnicity or cultural-linguistic backgrounds. The process entailed training teacher educators to independently evaluate the PACT assessment materials and determine through a consensus process whether sources of bias were evident. Panel participants used a structured reporting form to record their responses to the PACT materials. The findings from this process were used to flag areas of potential bias which informed subsequent revisions of the Teaching Event 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 2003-04 pilot, there were no significant differences in scores by race/ethnicity of candidates, percent of ELL students in candidates' classrooms, grade level taught (elementary versus secondary), academic achievement level of candidates' students, and months of previous paid teaching experience. There were statistically significant differences between male and female candidates (with females scoring higher) and between candidates teaching in schools in different socio-economic contexts (with candidates in suburban schools scoring slightly higher than those in urban or inner-city schools). PACT will continue to monitor and re-examine the scorer training process and the design of the TE assessments to uncover any potential sources of bias that may exist due to varying socio-economic contexts.

Construct Validity. Factor analyses were conducted on the score data from the 2003-04 pilot years. Two factors emerged from the scores: factor one was comprised of Planning, Instruction, and Academic Language items, and factor two was comprised of Assessment and Reflection items. These results suggest that the assessment tasks (Planning, Instruction, Assessment, and Reflection) are meaningful constructs that represent significant domains of teaching skill.

Criterion-Related Concurrent Validity. Using the final passing standard resulting from the standard setting process, we were able to compare candidates' pass/fail status (based on their Teaching Event scores in comparison to the passing standard) with the holistic judgments made by scorers as to whether they should or should not be recommended for a credential based on their reading of their Teaching Events. The sum of exact matches in the pass/fail decision (3.8% Not Recommended AND Fail + 85.0% Recommended AND Pass) is 88.8%. This level of consistency (exact matches) is very high and provides evidence of the validity of the assessment passing standard to evaluate candidate performance.

Score Consistency and Reliability. In these analyses, we examined the consistency across pairs of scores by computing consensus estimates within each subject area. These analyses are consistent with those documented in the ETS TPA's score analysis report.

During the 2003-04 pilot year, we examined the level of agreement between campus-assigned scores and Audit scores for the same Teaching Events. We found that 91% of score pairs were exact matches or within one point.

Assessor Reliability Estimate. Inter-rater reliability was also calculated using the Spearman Brown Prophecy reliability statistic in order to maintain consistency with the ETS TPA's score analysis report. For the 2003-04 pilot year, the overall inter-rater reliability for all rubrics across tasks is 0.880.

Standard Setting. Based on standard setting models described by Haertel (2002) and Haertel & Lorie (2000), as well as the process used by the National Board for Professional Teaching Standards (Phillips, 1986), we utilized a three-stage process that first convened a panel of teacher educators familiar with the PACT scoring process to formulate initial recommendations for the passing standard on the Teaching Event. The second stage called for a confirmatory group from the PACT programs to review those initial recommendations and to decide on a set of passing standards (along with a cut-score model) that would be submitted to all participating programs for review and approval. Based on consensus agreement from all PACT consortium members, the standard setting process resulted in a passing standard that valued all five components (Planning, Instruction, Assessment, Reflection, Academic Language) equally. The final step involved the gathering of policymakers (deans of education and directors of teacher education) across the PACT Consortium to review and approve a final set of passing standards. This step was completed in January 2007.

The **Final Passing Standard** that was ultimately selected by program directors and deans in January 2007 is below.

Candidates pass the Teaching Event if they pass ALL FIVE rubric categories (Planning, Instruction, Assessment, Reflection, and Academic Language) AND have no more than 3 scores of "1" across the tasks.

The cut score for each category is as follows: 1.66 for Planning (1 out of 3 scores can be a "1"); 1.5 in Instruction, Assessment, Reflection, and Academic Language (1 out of 2 scores can be a "1")

In other words, there are two ways a candidate could fail the Teaching Event

- 1) Fail one or more rubric categories
- 2) Have more than 3 failing rubric scores of "1" in categories across the Teaching Event

The overall pass rate for 2003-04 pilot year Teaching Events, using a slightly modified version of this passing standard (to reflect slight differences in the rubric structures), was 85.4%.

Introduction

In 1998, California elected to require teacher preparation programs to use standardized performance assessments in making credentialing decisions, along with other measures (e.g., coursework, practicum observations). California contracted with the Educational Testing Service (ETS) to develop such an instrument, but gave teacher education institutions the option of using a different instrument if it met the CTC's Assessment Quality Standards.) A coalition of California institutions of higher education formed PACT (Performance Assessment for California Teachers) to develop such an alternative assessment method. The development of the PACT has been funded by participating institutions, the University of California Office of the President (UCOP), the Flora and Sally Hewlett Family Foundation, the Hewlett Foundation, and the Morgan Family Foundation. A key motivation for the PACT consortium is to develop an integrated set of rigorous, transparent, subject-specific, standards-based certification assessment instruments that are consistent with the curricular and professional commitments of the member institutions. The goal of the PACT consortium is to strengthen the quality of teacher preparation by using curriculum-embedded assessment instrument(s) developed by each member institution in combination with a standardized teaching performance assessment to recommend licensure for prospective teachers. To achieve this purpose the PACT TE alternative was conceptualized as an "evidence-based system" that uses multiple sources of data (teacher plans, teacher artifacts, student work samples, video clips of teaching and personal reflections and commentaries) that are organized around four categories of teaching (Planning, Instruction, Assessment and Reflection).

The PACT consortium was initially composed of the following 12 universities: UC Berkeley, UCLA, UC San Diego, UC Santa Cruz, UC Santa Barbara, UC Riverside, UC Davis, UC 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, California State University-Sacramento, the San Diego City Schools Intern Program, and the University of Southern California. In the summer of 2005, an additional campus, CSU-Dominguez Hills, joined the consortium; in June 2006, the University of the Pacific also joined; and in October 2006, CSU-Northridge joined the PACT consortium.) As of February 2007, the PACT consortium included 27 universities. New members include: Sonoma State University, CSU-Channel Islands, CSU-Monterey Bay, University of San Diego, Pepperdine University, Notre Dame de Namur University, Saint Mary's College of California, and Holy Names University. The proportion of newly licensed teachers in the state credentialed by PACT consortium institutions is 29.2% (8,017 teachers out of 27,467 newly credentialed teachers in California, based on the 2004-05 Title II report).

Assessment Design

The Performance Assessment for California Teachers (PACT) project focuses on two assessment strategies: (1) the formative development of prospective teachers through "Embedded Signature Assessments" (ESAs) that occur throughout teacher preparation, and (2) a summative assessment of teaching knowledge and skills during student teaching (called the "Teaching Event"). In 2002, 12 colleges and universities began collaborating

to identify and share exemplary curriculum-embedded assessments across programs. These embedded assessments include child case studies, planning instructional units, analyses of student work, and observations of student teaching. Through a website these assessments are being shared across the PACT institutions. These ESAs are used to monitor candidates' progress toward meeting the TPEs and are used formatively to provide feedback to teacher candidates prior to completion of the summative Teaching Event.² The Teaching Event developed by PACT was purposefully designed to fully address the state teaching standards - Teaching Performance Expectations (TPEs) - and the assessment standards needed to meet CCTC licensure requirements.

The following section addresses the first part of the CCTC's Assessment Quality Standard 19 (a): *"The Teaching Performance Assessment includes complex pedagogical assessment tasks to prompt aspects of candidate performance that measures the TPEs."* **The second part of this element is satisfied by the section on Content Validity, found on pages 25-27 of this report.**

The Teaching Events are subject-specific assessments linked to the California content standards for students, and are integrated across four tasks: planning, instruction, assessment, and reflection (PIAR) (with a focus on Academic Language³ embedded across the tasks). For each Teaching Event, candidates must plan and teach a learning segment comprised of 3 to 5 hours of instruction (i.e., an instructional unit or part of a unit), videotape and analyze their instruction, analyze student learning, and reflect on their practice. The Teaching Events are designed to measure and promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions and reflecting on practice. By probing candidate thinking about student learning, completing the assessments provide important opportunities for mentoring and self-reflection. Additional benefits include focusing attention on the academic language 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.

In Task A ("Planning Curriculum, Assessment, and Instruction"), teacher candidates begin by describing the instructional context in which they will be teaching the learning segment for the Teaching Event. In order for raters to understand their teaching decisions, candidates are asked to write a commentary of about two pages that describes key characteristics of the class that affect the planning and teaching of the learning segment, such as characteristics of students in the class, the curriculum, and instructional

² We are currently planning a series of research studies to support the validity and reliability of ESAs so that they may eventually be used as part of the summative credentialing decision in combination with the Teaching Events.

³ Academic Language is defined as "the language needed by students to understand and communicate in the academic disciplines." Academic language includes such things as specialized vocabulary, conventional text structures within a field (e.g., essays, lab reports) and other language-related activities typical of classrooms, (e.g., expressing disagreement, discussing an issue, asking for clarification). Academic language includes both productive and receptive modalities. For additional examples of academic language across the content areas, see the Academic Language rubrics. See also pp.43-45 of **Appendix F. Handbook for Implementing PACT Scorer Training** for additional explanations of Academic Language.

context, including any constraints on their teaching. Candidates also complete an instructional context form in which they report the number of students in the class, the grade level of the class or any specialized features, the number of special needs and English learners, the title of the textbook used (if any), and the number of available computers in the class and school. Candidates then provide an overview of their planned learning segment spanning 3-5 hours of instruction, lesson plans for each lesson, assignments and other instructional materials for the learning segment.

In Task B (“Implementing Instruction”), candidates videotape one or more of their lessons from the learning segment, select up to 15-20 minute clips of the video (based on criteria set for each content area) and write a commentary on the unedited video clip(s) they have selected. In their commentary, candidates describe the context of the video clip (what happened before and after the clip); routines or working structures seen in the clip and how students were prepared for these routines; the ways in which the candidate engaged students with the lesson content; strategies used to address specific individual learning needs; and any language supports provided to students to understand the content or academic language.

In Task C (“Assessing Student Learning”), candidates collect and analyze student work from the learning segment. In the whole class learning commentary, candidates are asked to provide a context for the assessment, including a rationale for selection and the conditions under which students completed it; summarize student learning across the whole class relative to the learning goals; and discuss what most students seem to have understood and any misunderstandings, confusions, or special needs. In addition, candidates propose next steps in instruction based on their analysis of student learning. In the individual student learning commentary, candidates select two students in the class (who represent different instructional challenges) to focus on in analyzing student learning over time. In this task, candidates collect and analyze three samples of each student’s work that reflect his or her growth or progress with respect to a central goal of your class. Candidates are also asked to describe the feedback provided to students on their work.

In Task D (“Analyzing and Reflecting on Teaching and Learning”), candidates are prompted to reflect daily on their lessons after each day of instruction. At the end of the learning segment, candidates are asked to reflect on what they learned from their teaching of the learning segment and to describe what they would do differently if they were to teach the same content to the same group of students. They are also prompted to explain how their proposed changes would improve the learning of their students and to cite specific evidence and theoretical perspectives and principles that inform their analyses.

Although Academic Language is not a task in the Teaching Event, it comprises an analytic category in the scoring rubrics. The Academic Language rubric is scored based on evidence drawn from all of the tasks. Teacher candidates are prompted in the Planning and Instruction tasks to describe how their lessons and instruction help to build students’ acquisition and development of Academic Language. For example, in Task A, candidates are prompted to describe the language demands of the learning and assessment

tasks that are likely to be challenging for their students. They are also asked to describe how they planned to support students in meeting those language demands. Task B asks candidates to describe any language supports they used to help students understand the content and/or academic language. Task C asks candidates to discuss the progress in learning over time for two students, one of which must be an English Learner or another student who is struggling with academic English. Reflection on the successes and problems in each lesson with respect to developing language proficiency is prompted in Task D.

The following section addresses the first part of the CCTC’s Assessment Quality Standard 19 (b): *“To preserve the validity and fairness of the assessment over time, the sponsor may need to develop and field-test new pedagogical assessment tasks and multi-level scoring scales to replace and strengthen prior ones.”* **The second part of this standard is addressed in the section on Construct Validity (page 30) and Criterion-Related Concurrent Validity (page 32).**

In 2002-03, PACT developed assessments in five certification areas: multiple subjects (elementary literacy and elementary math), English/language arts, mathematics, history/social science, and science. (See **Appendix D: Timeline of PACT Activities & Participants**, pp.66-76) for a summary of assessment development activities and participants.) During the 2003-04 pilot year, PACT revised the Teaching Events in these areas in consideration of feedback received from candidates and teacher educators who had piloted the assessment in the first year, as well as findings from the bias and fairness review of the 2002-03 pilot handbooks and rubrics. Minor revisions to the handbook and rubrics have been made each year in response to field testing results. Current versions of the Teaching Event Handbook and Rubrics do not differ very much from the 2003-04 version. (See **Appendix A** for an overview the Elementary Literacy Teaching Event and sample rubrics for the 2003-04 pilot.) The consortium has created assessments in several additional areas to include a full complement of certification areas offered by members of the PACT consortium, including world language, art, music, and physical education. Teaching Events for B-CLAD and special education teacher candidates are also planned for development, validation, and implementation by July 2008.

The purpose of this technical report is to describe the results of the 2003-04 pilot test of the Performance Assessment for California Teachers (PACT) Teaching Events and summarizes the assessment reliability and validity evidence conducted. A description of score analyses and summaries of results follow a brief description of the scoring system including a summary of scorer training.

The following section addresses the first part of the CCTC’s Assessment Quality Standard 20(g): *“The sponsor ensures equivalent scoring across successive administrations of the assessment and between the Commission’s prototype and local assessments by: using marker performances to facilitate the training of first-time assessors and the further training of continuing assessors...”*

Scoring System

The scoring system includes rubrics, benchmarks, scorer recruitment and training, and a scoring process. A common scoring system template is used in each subject area, but benchmarks, scorers, and some scoring rubrics are subject-specific. In the 2003-04 pilot year, each set of rubrics was comprised of 13 separate “Guiding Questions” and rubrics that fall into five major categories (Planning included 5 rubrics, Instruction included 2 rubrics, Assessment included 3 rubrics, Reflection included 2 rubrics, and Academic Language included 1 rubric.) Each rubric is scored on a 1-4 scale. (See **Appendix A**, pp.52-57 for sample rubrics.) Rubric level 1 represents a performance that is developing but does not yet meet the passing standard of performance. It does not, however, represent a complete absence of competency. Level 2 represents a performance that meets the passing standard of performance. Level 3 represents a performance that clearly exceeds the passing standard with many strengths. Level 4 represents a performance that is highly competent and would generally be exceptional for a beginning teacher.

Benchmarking. In the 2003-04 pilot year, scores from the first pilot year were used to identify Teaching Events as potential benchmarks for rubric levels 1, 2, and the higher levels (3 or 4) in each content area. Pairs of benchmarking participants were recruited from the first pilot year’s trainers, scorers, and nominated faculty from PACT institutions. Each benchmarking team was given two potential benchmarks at different levels. The previous scores and scoring levels were not shared with the participants. They scored the Teaching Events, and made a recommendation as to the suitability of each Teaching Event for use as a benchmark. Each pair of benchmarking participants had discussions to reach consensus on which TEs to select for each rubric level benchmark, and agreed on an assignment for formally writing up the evidence for each benchmark. The rubric language and written documentation were reviewed by Kendyll Stansbury, a PACT assessment specialist, and Raymond Pecheone, the Project Director. Ultimately, three benchmarks (representing not passing, passing, and a strong performance) for each content area were selected for use in scorer training.

Scorer recruitment. In both pilot years, scorers were recruited by subject area and included faculty, supervisors, cooperating teachers from across the state and National Board Certified teachers. To be eligible to score, scorers had to meet the following criteria:

- At least three years experience with teaching K-12 students or supervising student teachers in the teaching credential area scored
- Experience providing feedback on the work of emergency permit teachers, pre-interns, interns, student teachers or beginning teachers (i.e., experience in assessing beginning teaching)
- Pedagogical content knowledge, content knowledge, and content standards in the teaching credential area scored
- Availability for the entire four-day scoring session

The following section addresses the CCTC's Assessment Quality Standard 19(c): *“The sponsor develops scoring scales and assessor training procedures that focus primarily on teaching performance and that minimizes the effects of candidate factors that are not clearly related to pedagogical competence, which may include (depending on the circumstances) factors such as personal attire, appearance, demeanor, speech patterns and accents that are not likely to affect student learning.”* **See also Appendix F. Handbook for Implementing PACT Scorer Training (pp.5-9, 29-31) for detail on anti-bias training.**

This section and Appendix F also address Assessment Quality Standard 20(c): *“The Teaching Performance Assessment system includes a comprehensive program to train assessors who will score candidate responses to the pedagogical assessment tasks. An assessor training pilot program demonstrates convincingly that prospective and continuing assessors gain a deep understanding of the TPEs, the pedagogical assessment tasks and the multi-level scoring scales. The training program includes task-based scoring trials in which an assessment trainer evaluates and certifies each assessor's scoring accuracy in relation to the scoring scales associated with the task. When new pedagogical tasks and scoring scales are incorporated into the assessment, the sponsor provides additional training to assessors, as needed.”*

Scorer training. Two scoring models have been piloted to determine the most feasible system that is also optimal for score reliability. For the first year pilot (2002-03), a centralized scoring model was utilized to get a better understanding of the training needs. An English-language arts trainer (Dr. Steven Athanases, UC-Davis) developed a training template based on Teaching Events from early completers. Regional lead trainers received training in this model and used Teaching Events in their content area to customize the training and apply the subject-specific rubrics. Centralized regional training was conducted at five sites - San Jose State, UCLA, Stanford University, UC-San Diego, and UC-Irvine. Each scoring session consisted of two days of training and two days of scoring. A total of 124 scorers participated in scoring the Teaching Events during the first pilot year. These participants are listed in **Appendix D. Timeline of PACT Activities & Participants.**

For the 2003-04 pilot, scoring was done within each local program based on a Trainer of Trainers model. Local programs sent subject-specific trainers to a one-and-one-half day Training of Trainers session at Stanford University or UCLA. (See **Appendix D. Timeline of PACT Activities & Participants**, pp.80-87 for a summary of 2003-04 assessment implementation activities and participants.) The local trainers then trained local scorers, either within a single program or in nearby programs. Scoring implementation strategies differed across programs, with some programs spreading the scoring over time, some doing it early, and others conducting the scoring within a centralized local scoring session. 69 local trainers participated in training scorers across campuses (these are listed in **Appendix D**, pp.83-85). After scoring, local programs were asked to identify a 20% random sample across content areas for rescoring at a centralized audit site. 48 scorers from across the PACT member institutions participated in the Score Audit Session at Stanford University.

Prior to reading benchmarks, scorers received an orientation to the scoring process. This covered the following topics:

- Task structure of Teaching Event (PIAR) and the task-based scoring procedure
- Collection of data on confidence in ratings and independent rating of candidate
- Relationship of TPEs to Guiding Questions
- Potential sources of bias
- Process for taking notes and documenting evidence
- Rubric Levels
- Preponderance of evidence approach for selecting score

Minimizing bias. In the training process, scorers are provided with specific training to identify potential sources of bias in scoring, including candidate characteristics irrelevant to their teaching competence such as their writing skill, the use of particular curricula, race/ethnicity, or other personal characteristics. Scorers are also asked to explicitly identify their own personal biases in order to minimize the role of factors irrelevant to teaching effectiveness in the scoring process. See **Appendix F. Handbook for Implementing PACT Scorer Training** (pp.5-9, 29-31) for detail on anti-bias training.

Benchmark review and calibration. During each year’s training, two or three benchmarks were read and debriefed, with increasingly independent reading and scoring during the training session. The focus of the training was to establish a common understanding (consensus) around the interpretation of the subject-specific benchmarks and to record evidence to support the benchmark ratings. At the end of day two of the audit scoring session, a pre-scored Teaching Event was scored independently by the scorers as a calibration assessment. Discrepancies were then discussed. For local scoring in the second year, scorers in most programs scored the same Teaching Event to assess consistency before independent scoring began. However, because the first two years of implementation were pilot years, none of the scorers who were trained was excluded from scoring. Therefore, we can expect greater reliability and consistency across scorers when a calibration process that screens out scorers is established and applied.

During the second year scoring audit, scorers independently scored TEs in their subject areas during days three and four of training. Trainers monitored the scoring process and provided technical assistance when needed. Selected TEs were double scored to assess scorer consistency. In the 2003-04 pilot year, about 9% were double-scored at local campuses and 23% were audited. For local scoring in the second year, some programs scored Teaching Events at a centralized site, and others used independent scoring scheduled at the scorers’ convenience.

The following section addresses the CCTC’s Assessment Quality Standard 20(d): *“In conjunction with the provisions of Standard 22, the sponsor plans and implements periodic evaluations of the assessor training program, which include systematic feedback from assessors and assessment trainers, and which lead to substantive improvements in the training as needed.”*

Each year, during the central Training of Trainers and scoring training at each local campus, feedback has been solicited from trainers and scorers using standard forms. This feedback has been used to inform changes in the Teaching Event handbooks, rubrics, and the training design. Changes in the training design have included the selection of new

benchmarks that are more appropriate representations of a score level and the inclusion of additional resources to support scorers' understanding of the meaning of each score level for each rubric (e.g., "Thinking Behind the Rubrics" document, found on pages 32-45 of **Appendix F. Handbook for Implementing PACT Scorer Training**). Training of trainers and local scorer training are conducted annually, using new benchmarks. Any changes in the pedagogical tasks and scoring scales are reflected in the new training materials.

Scoring Process. To score the TE, scorers use a task-based scoring model. The score path follows the design of the portfolio in a progressive or (sequential) manner as represented by the following process: (1) scorers read and record evidence related to the Context/Planning category of the TE and score the rubrics for the associated Guiding Questions (GQs)⁴; (2) scorers view the video clips, read the commentary and score the GQs related to the Instruction category; (3) scorers evaluate representative student work samples for the whole class as well as the commentary on student learning, and then score the GQs associated with the Assessment category; and (4) scorers read the candidate commentaries/reflections and score the GQs related to the Reflection category. For the 2003-04 pilot year, scorers rated a single holistic Academic Language rubric that relied on evidence gathered across the tasks.

A systematic process of evidence gathering supports the scorers' rubric ratings. The following procedural steps are employed: (1) scorers independently take notes as they read through each portfolio task (PIAR); (2) scorers independently read all documents that have been provided by the candidate to illustrate their teaching - lesson plans, assignments, reflective commentaries - as well as student work samples; (3) scorers independently review their raw notes and construct evidence to support their rubric ratings on each of the GQs; and (4) scorers record their scores for each GQ on a standardized form. For the 2003-04 pilot, scorers were given the option to write a summary of the supporting evidence.

As a result of this process, a detailed score profile is generated that provides information at the item level (GQ) and at the analytic category level – Planning, Instruction, Assessment, Reflection, and Academic Language. Individual score profiles may be used by candidates to develop an Individual Induction Plan for use in a professional induction program (as stated in the CCTC's Program Standard 21(e), formerly 23(e)). To develop a meaningful task based score, the raw total score and PIAR sub-scores were divided by the number of guiding questions within each analytic category, resulting in an average score ("mean item score") ranging from 1 to 4, reflecting the rubric scale. These "mean item scores" provide for greater interpretability of the results. Aggregated category scores for all candidates within a program may be used as a basis for internal and external reviews of the teacher preparation program for the purpose of program improvement, as required by the CCTC's Program Standard 21(g), formerly 23(g).

⁴ Guiding questions represent the big ideas around teaching measured in each of the rubrics, e.g., GQ1 (Planning): "How does the instructional design make the curriculum accessible to the students in the class?"

The following section addresses the CCTC's Assessment Quality Standard 20(e): *"The program sponsor requests approval of a detailed plan for the scoring of selected assessment tasks by two trained assessors for the purpose of evaluating the reliability of scorers during field-testing and operational administration of the assessment. The subsequent assignment of one or two assessors to each assessment task is based on a cautious interpretation of the ongoing evaluation findings."*

This section also addresses the second part of the CCTC's Assessment Quality Standard 20(g): *The sponsor ensures equivalent scoring across successive administrations of the assessment and between the Commission's prototype and local assessments by... monitoring and recalibrating local scoring through third-party reviews of scores that have been assigned to candidate responses; and periodically studying proficiency levels reflected in the adopted passing standard."*

This section also addresses the third part of the CCTC's Assessment Quality Standard 20(h): *"The sponsor demonstrates that the assessment procedures, taken as a whole, maximize the accurate determination of each candidate's overall pass-fail status on the assessment."*

Proposed Scoring System Design

The system that supports the scoring of the Teaching Event is as follows:

Each local campus will have a group of subject-specific trainers or will enter into agreements with other programs in the PACT consortium to share one or more trainers. The trainers will be prepared through a rigorous Training of Trainers program that will be repeated annually. Trainers will need to reach a calibration standard in order to be eligible to work as a trainer. The trainers will then assume a set of responsibilities that include training, calibrating, and supervising local scorers.

All Teaching Events will be independently scored at least once by trained and calibrated scorers at each local campus. All Teaching Events with scores that do not meet the established passing standard and borderline scores (those just above the passing standard) will be scored by two trained local scorers, and the evidence reviewed by the chief trainer, to ensure the reliability of the scores. In addition, as Teaching Events are scored, a randomly selected stratified sample of 10% of TEs from across the score levels (2s and 3/4s) and across scorers will be double-scored. If the scores given by two different scorers conflict by two or more rubric levels on any rubrics, or the differences in scores result in different pass/fail outcomes, the trainer in the specific content area will also score the Teaching Event to resolve discrepancies. Trainers will monitor the double scoring by examining the scores for Teaching Events that were double-scored and conducting "read behinds" for scores that are widely discrepant. The chief trainer will identify scorers who are drifting and will work with them to again achieve calibration by discussing the discrepant scores and helping the scorers to understand the differences between levels on rubrics that appear to be problematic for the scorers.

To ensure that scoring is calibrated across campuses, the trainers will participate in a central audit of all failing Teaching Events and a randomly selected stratified sample of 10% of Teaching Events from across the score levels (2s and 3/4s) from across content areas and across all campuses. Audited Teaching Events that have large score

discrepancies (2 or more points) from local scores will be rescored by other trainers as part of a moderation process to ensure consistency. If there is sufficient evidence that local campuses have unreliable scores, the scoring process at those campuses will be monitored closely in the following year by a trainer external to the campus. If the local campus scores continue to have large discrepancies with audit scores in the second year, external trainers will be sent to conduct the local campus training and supervise scoring.

Every third year, a central standardized scoring model will be used to provide another check on the consistency of training and the scoring process and the reliability and validity of scores. Under this model, local scorers will be convened at central scoring sites within a region to be trained and calibrated, and to score Teaching Events.

Remediation process for candidates whose Teaching Events do not meet the passing standard. If candidates fail the Teaching Event (See Final Passing Standard on p. 47) because they fail more than one task, OR have more than 3 “1”s across tasks, an entirely new Teaching Event must be re-taught and re-submitted. However, candidates who fail the Teaching Event because they failed only one task of the Teaching Event have the opportunity to resubmit specific individual tasks for a higher score. With the exception of the Reflection task, resubmitting a task involves more than simply re-writing/revising the commentary for an individual task. The chart below shows what would need to be resubmitted for each task that is failed.

Task Failed	Components to be resubmitted
Planning	Instructional context task; New series of lesson plans and instructional materials on a new topic; Planning commentary
Instruction	Instructional context task; New video clips; New lesson plans for the lessons from which the video clips are drawn; Instruction commentary
Assessment	Instructional context task; New student work samples; Assessment commentary
Reflection	Revision of reflection commentary for previously taught Teaching Event; Daily reflections cannot be revised.*
Academic Language	Instructional context task; New Planning Task + New Instruction Task (See above for components to be resubmitted)

* Guiding Question 8 (Reflection 1) on the current version of the rubrics are based on the Daily Reflections exclusively, and since Daily Reflections depend on teaching the learning segment, the score for this guiding question cannot be remediated.

The following section addresses the CCTC’s Assessment Quality Standard 20(i): *“The sponsor’s assessment design includes an appeal procedure for candidates who do not pass the assessment, including an equitable process for rescoring of evidence already submitted by an appellant candidate in the program.”*

Appeals procedure. Candidates whose Teaching Events do not meet the passing standard and who choose not to remediate the score by resubmitting a task or an entire Teaching Event will have the right to appeal the failing score. As noted above, all Teaching Events not meeting the passing standard will have already been scored at least

twice by trained scorers, and the evidence reviewed by the chief trainer (a “read-behind”), to ensure the reliability of scores. If the original double scores were conflicting, then the chief trainer will have independently scored the Teaching Event a third time to adjudicate the scores. If a candidate appeals the failing score, an investigation of the scorer training and scoring procedures at the local campus will be triggered. If the investigation finds that the scorer training process at a local campus or scoring procedures were not in accordance with the scoring system as designed, the candidate then has the right to ask for a re-scoring of the Teaching Event by trained scorers external to the local program. The re-scoring of the Teaching Event must occur within a month of the original appeal to allow the candidate time to re-submit a task or an entire Teaching Event should the re-scoring of the Teaching Event results in a failing score.

The following section addresses the CCTC’s Assessment Quality Standard 19(h): *“In designing assessment administration procedures, the sponsor includes administrative accommodations that preserve assessment validity while addressing issues of access for candidates with disabilities.”*

Access for Candidates with Disabilities. The term "candidates with disabilities", as it is used in this report, refers to teacher candidates who are eligible for services under the Individuals with Disabilities Education Act (IDEA) as well as candidates who are covered under Section 504 of the Rehabilitation Act of 1973 (Section 504) and Title II of the Americans with Disabilities Act (ADA). Under IDEA, a student is eligible for services if the student has one of the covered impairments and because of that impairment needs special education and related services. Under Section 504 and Title II, the student is covered if the student has a physical or mental impairment which substantially limits one or more major life activities such as learning.

In accordance with required educational accommodations as outlined in the federal Individuals with Disabilities Education Act,

Students with disabilities must be provided with appropriate accommodations when necessary to enable participation in the assessments. Assessment accommodations include changes in the way assessment items are presented, changes in the way a student may respond, changes in the timing or scheduling of an assessment, and changes in the setting that are used to provide an equal footing for students with disabilities who need the accommodations. Assessment accommodations help students show what they know without being placed at a disadvantage by their disability. (United States Department of Education, 2003)

Following the guidelines utilized by the Educational Testing Service (2007), each PACT consortium member will be expected to provide reasonable accommodations for candidates with documented disabilities, recognized under the Americans with Disabilities Act, which mandates that test accommodations be individualized, meaning that no single type of test accommodation may be adequate or appropriate for all individuals with any given type of disability. PACT institutions serving candidates with documented disabilities will be required to establish a clear policy for the accommodation

of disabled candidates completing the PACT Teaching Event. The assessment accommodations provided should be consistent with the accommodations that disabled candidates were entitled to have during the course of the credential program.

Some accommodations that may be considered include:

- An extension on deadlines for submission
- Reader
- Recorder/writer of answers
- Sign language interpreter (for spoken directions provided in teacher education classes)
- Technical assistance with videotaping for the Instruction task
- Braille
- Large print handbooks
- Large-print rubrics
- Audio recording
- Audio recording with large-print figure supplement
- Audio recording with raised-line (tactile) figure supplement

For Teaching Events completed on computer-based platforms:

- Selectable background and foreground colors
- Alternate handbook formats: Audio recording; Braille; Large print
- Kensington Trackball mouse
- HeadMaster™ Plus mouse
- IntelliKeys® keyboard
- Keyboard with touchpad
- Magnifying Text on Computer Screens

Since the accommodations listed above are minor and do not significantly alter what is measured, the score report for the candidate's Teaching Event would contain no indication of whether the assessment was completed with accommodations.

The following section addresses the CCTC's Assessment Quality Standard 20(a): *"In relation to key aspects of the major domains of the TPEs, the pedagogical assessment tasks and the associated directions to candidates are designed to yield enough evidence for an overall judgment of each candidate's pedagogical qualifications for a Preliminary Teaching Credential. The program sponsor will document sufficiency of candidate performance evidence through thorough field-testing of pedagogical tasks, scoring scales, and directions to candidates."* See also the section on **Criterion-Related Concurrent Validity, page 32.**

This section also addresses Assessment Quality Standard 20(b): *"Pedagogical assessment tasks and scoring scales are extensively field-tested in practice before being used operationally in the Teaching Performance Assessment. The sponsor of the program evaluates the field-test results thoroughly and documents the field-test design, participation, methods, results and interpretation."*

2003-04 Pilot Year– Score Summary

To test the utility and viability of the PACT Teaching Event within and across subject areas, the assessment was piloted across 13 PACT programs in 2003-04. The number of pilot candidates varied across institutions, with each PACT institution purposely selecting particular subject areas or cohorts of students to participate in each year’s pilot. In the 2003-04 pilot, teacher candidates produced Teaching Events in elementary (either literacy and mathematics) and secondary subjects (English/language arts, history/social studies, mathematics, and science).

Pilot Sample. During the 2003-04 pilot year, approximately 700 Teaching Events were completed and submitted, and scores for 625 Teaching Events were submitted by campuses (not all Teaching Events were scored in time for the score submission deadline). Of those 625 Teaching Events, 146 (23%) were audited at the Score Audit session in June 2004 held at Stanford University. Additionally, some campuses double-scored a small percentage of their candidates’ Teaching Events, although the process used for double scoring varied across campuses, with some utilizing paired scoring methods (in which raters were able to talk about their scores or came to a consensus about the scores). A small percentage of Teaching Events were also double or triple-scored at the Score Audit session. In the absence of an official passing standard, a Teaching Event was rescored when it had a number of rubric item scores that had three or more scores of “1” or when two raters’ scores of the same Teaching Event were discrepant by more than one point.

As **Table 1** below indicates, the subject area with the highest number of Teaching Events submitted and scored in the 2003-04 pilot was Elementary Literacy, while the number of Teaching Events scored in Elementary Mathematics was half of those scored in EL. During the second pilot year, many institutions allowed candidates to submit a single Teaching Event in either Elementary Literacy or Elementary Mathematics.⁵ Tasks were also submitted in the following Single Subject areas: English Language Arts, Mathematics, History-Social Science, and Science.

⁵ In the piloting phase, it has been our policy to allow multiple subjects candidates to choose to complete either the Elementary Literacy or Elementary Mathematics Teaching Event. When PACT is operational (after July 2008), multiple subject candidates will complete Teaching Event tasks in both literacy and mathematics. Specifically, PACT consortium members will continue to offer the candidates the choice of completing a Teaching Event in either literacy or mathematics. However, multiple subject candidates will also complete a standardized performance task in the content area not covered by the selected Teaching Event (Elementary Literacy or Elementary Mathematics) to demonstrate their competency in teaching both of these core content areas. The additional performance tasks will be drawn from the existing tasks within the Teaching Event (e.g., Planning, Instruction, or Assessment) and would be scored using the same training, rubrics and benchmarks used to score the Teaching Events. This task-based design of the multiple subjects PACT assessment is consistent with the design of the California TPA.

Table 1. Total Teaching Events Scored by Subject Area (2003-04 Pilot Year)

Subject Area	#TEs Scored at Campuses*	# Double-scored at Campuses	# Scored at Audit Session**
ELEMENTARY LITERACY (EL)	224 (35.8%)	19 (34.5%)	79 (35.9%)
ELEMENTARY MATH (EM)	109 (17.4%)	7 (12.7%)	47 (21.4%)
ENGLISH LANGUAGE ARTS (ELA)	110 (17.6%)	7 (12.7%)	31 (14.1%)
MATHEMATICS (MTH)	50 (8.0%)	8 (14.5%)	22 (10.0%)
HISTORY-SOCIAL SCIENCE (HSS)	60 (9.6%)	7 (12.7%)	14 (6.4%)
SCIENCE (SCI)	72 (11.5%)	7 (12.7%)	27 (12.3%)
Total	625 (100.0%)	55 (100.0%)	220 (100.0%)

* Represents the total number of Teaching Events scored across all campuses and includes those that were double-scored and scored at the Audit Session. **Audit scores include double scores and multiple scores for Teaching Events used for calibration. Actual number of Teaching Events scored at the Audit Session was 146.

Teacher Candidate Sample. Preservice teachers who completed Teaching Events in the 2003-04 pilot year were enrolled in 11 different teacher preparation programs across the state: Information on the demographic backgrounds of candidates completing Teaching Events was collected through an online survey that was completed after submission of the Teaching Event. 287 teacher candidates with scored Teaching Events completed the online survey in 2003-04, for a response rate of 46%. The demographic information of survey respondents is summarized below in **Table 2**.

Table 2. Demographic Characteristics – PACT Candidate Survey (2003-04)

	Piloting candidates who completed the online PACT Candidate Survey and whose Teaching Events were scored	
	Number	Percent
Gender		
Female	231	80.5
Male	54	18.8
Missing	2	.7
Race or ethnicity		
African American	15	5.2
American Indian/Alaskan Native	1	.3
Asian	62	21.6
Filipino	7	2.4
Hispanic/Latino/Chicano	67	23.3
White	125	43.6
Other	6	2.1
Missing	4	1.4
Type of teacher candidate		
Multiple Subject	170	59.2
Single Subject	117	40.8
Practicum Type		
Student Teacher	227	79.1
Intern	57	19.9
Emergency Permit Teacher	3	1.0
Subject area		
EL	120	41.8
EM	50	17.4
ELA	47	16.4
MTH	31	10.8
HSS	11	3.8
SCI	28	9.8

Score Profiles (2003-04 Pilot Year). This section provides an overview of scores across subject areas, summarizing how well teacher candidates did on the Teaching Event across all of the PACT institutions that participated in the 2003-04 pilot year. For Teaching Events that were scored multiple times (double scored, scored as a calibration Teaching Event, or rescored for adjudication), the average of these scores was calculated to obtain average individual Teaching Event scores. The number of guiding questions in the score rubrics for each subject area was 13. The guiding questions were conceptually equivalent across subject areas but tailored with subject-specific language. Total scores are the sum of scores across the 13 rubrics. Subscores refer to the sum of scores within the analytic categories (Planning, Instruction, Assessment, Reflection, Academic Language).

Because it is difficult to assess the meaning of the total scores and subscores by themselves, they were converted to “mean item scores,” which refer to the total score or subscore divided by the number of guiding questions in each analytic category. Since each guiding question was scored on a scale of 1 to 4, each mean item score 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. Rubric level 1 represents a

performance that is developing but does not yet meet the passing standard of performance. Level 2 represents a performance that meets the passing standard of performance. Level 3 represents a performance that clearly exceeds the passing standard with many strengths. Level 4 represents a performance that is highly competent and would generally be exceptional for a beginning teacher. Thus, in the following tables, when a mean item score falls between 2 and 3, this indicates that on average, candidates have received a majority of 2s and 3s and have, on average, met the standards for minimally competent performance as a whole or on each of the tasks.

Table 3. Mean Item Scores (MIS) by Subject Area (2003-04 Pilot Year)

	Total MIS (Std. Dev.)	Planning MIS (Std. Dev.)	Instruction MIS (Std. Dev.)	Assessment MIS (Std. Dev.)	Reflection MIS (Std. Dev.)	Academic Language (Std. Dev.)	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

Note: **Audit scores were excluded from this score 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.

EL – Elementary Literacy

MTH – Secondary Mathematics

EM – Elementary Mathematics

HSS – Secondary History-Social Science

ELA – Secondary English Language Arts

SCI – Secondary Science

Scanning the mean item scores across the analytic categories, it is apparent that teacher candidates across most subject areas scored highest on the Planning analytic category. The total mean item scores indicate that the average performance across subject areas met the minimum standard of performance both within score areas and overall. They also show some variability across the score range across subject areas.

The mean item scores for each of the rubrics in **Table 4** below allow us to examine the particular rubrics on which candidates performed at higher and lower levels. (See **Appendix A** for the complete guiding questions for each rubric.) Candidates scored highest in the Planning and Instruction items. Relative to overall scores in the other analytic categories, the scores for Reflection were the lowest. The average score differences across items were not extremely large, however, ranging from 2.47 to 2.89.

The pilot sample above formed the basis for conducting the reliability and validity studies that were designed to address the state-adopted Assessment Quality Standards. The following section summarizes the validity argument that was used to frame and guide the research studies that were implemented or are being planned to validate the TE.

Table 4. Descriptives –Rubric Scores (2003-04 Pilot Year)

Rubrics	N	Mean	Std Dev
Planning			
Access to curriculum	625	2.89	.801
Coherent instructional design	623	2.77	.827
Balanced instructional design	623	2.71	.805
Student needs and characteristics	624	2.62	.817
Assessment alignment	623	2.89	.801
Instruction			
Engagement	617	2.77	.827
Monitoring learning	612	2.71	.805
Assessment			
Whole class learning	621	2.62	.817
Individual learning progress	621	2.62	.794
Feedback	621	2.53	.808
Reflection			
Focus of reflections	622	2.50	.796
Teaching and learning	623	2.47	.891
Academic Language			
Academic language	617	2.56	.938

Note: Audit scores were excluded from this score summary.

Summary of Validity and Reliability Analyses for the Teaching Event

The *Standards for Educational and Psychological Testing* (1999) published by the American Education Research Association, American Psychological Association, and National Council Measurement in Education, define validation as “developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use” (p.9). Establishing validity for the PACT TE at the most basic level focuses on the “use” to which the instrument is put, not on the TE itself (i.e., the interpretation of a score is the subject of validation, not the assessment itself.) In essence, validating the Teaching Event entails collecting evidence for the conclusions reached about the prospective teacher’s knowledge or skills (competence) in relationship to the TPEs. Because of the high-stakes implications of this evidence (initial teacher licensure), as required by SB 2042, it is imperative that we examine the extent to which candidate performance on the TE can accurately differentiate between effective candidates (prospective teachers who meet or exceed TPE standards) and ineffective candidates (prospective teachers who do not meet minimum TPE standards).

The next section highlights results from the studies that have been conducted to date to determine the validity and reliability of the Teaching Event.

The following section meets the second part of the CCTC’s Assessment Quality Standard 19(a):

“Each task is substantively related to two or more major domains of the TPEs. For use in judging candidate-generated responses to each pedagogical task, the assessment also includes multi-level scoring scales that are clearly related to the same TPEs that the task measures. Each task and its associated scales measure two or more TPEs. Collectively, the tasks and scales in the assessment address key aspects of the six major domains of the TPEs. The sponsor of the professional teacher preparation program documents the relationships between TPEs, tasks and scales.”

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 has been the primary source of validity used to support teacher licensure assessment (AERA, APA & NCME, 1999; Wilkerson & Lange, 2003). In the process of establishing content validity, the assessment developer frames the assessment around the knowledge and skills that are represented in the standards (e.g., the California Teaching Performance Expectations). Therefore, the structure, substance and representativeness of the Teaching Event tasks were examined in relationship to the TPE domains. To establish the connection between the TPEs and the TE, the specific elements of the TE (tasks, prompts, GQs and rubrics) were mapped to the TPE domains. Another key concern was to evaluate the degree of coverage of the PACT assessment in relationship to the skills and abilities that define the 12 TPE elements.⁶ (See **Table 5** below for a list of the TPE elements.) This analysis provides an estimate of the comprehensiveness or proportionality of the assessment tasks to the six domains of knowledge or skills that define competent practice embodied in the TPEs.

Methodology. Evidence of the content validity of the PACT assessments was based on the expert judgment of teacher educators in the PACT consortium. Teacher educators and other university faculty involved in designing the Teaching Event tasks and rubrics were asked to judge the extent to which the content of the PACT assessments was an authentic representation of important dimensions of teaching as represented by the TPEs and whether the subject-specific aspects of the TEs measured key aspects of each content discipline. The judgment of experts serves to confirm the connection between the tasks that are represented in the TEs and the elements of the TPEs that the tasks were specifically designed to measure.

Four groups contributed to establishing the content validity of PACT: subject specific assessment design teams, program directors, university faculty and the PACT leadership team. (See **Appendix D**, pp.66-76 for names of those who participated in designing the PACT assessments and members of the aforementioned groups.) Within each subject area, a group of teacher educators representing different campuses was charged with the design and development of the PACT assessment. In carrying out this work, the assessment developers evaluated (a) the relevance of each task to the student teaching experiences of the prospective teachers; (b) the consistency of each task with the K-12

⁶ In the 2003-04 TE handbooks and rubrics we did not address TPE 12: Professional, Legal, and Ethical Obligations because it is better assessed at the program level.

academic content standards; and (c) the authenticity of the work samples collected in the TE in relationship to activities that a prospective teacher might encounter on the job.

In addition, the PACT leadership charted the alignment of the PACT TE with the TPEs across all subject matters. Overall, the findings across all 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. (See **Appendix B. Relationship of TPEs to Guiding Question Rubrics.**) An analysis of the representation of the specific elements of the TPEs (taking each individual sentence as an element) to the Teaching Event was conducted by Jones (2005) (See **Appendix E**). A data base was constructed that consisted of the 180 individual declarative sentences with the set of TPEs. The Teaching Event assessment questions (and, if new ideas were introduced, subquestions) and rubrics were also entered into the data base. The two sets of data were then cross-referenced, and associations between the Teaching Event elements and the TPE elements were identified. Each association was then examined to determine whether or not the same concept was represented in the Teaching Event and the TPE element. If not, the association was removed from the frequency count. A series of tables and charts were then constructed to illustrate the representation of the TPEs in the Teaching Event. **Table 5** below shows the representation and distribution of the TPE elements using this methodology. Jones found that every TPE except TPE 12 (Professional, Legal and Ethical Obligations) was represented in the Teaching Event.

Table 5. TPE Elements Represented in the Teaching Event

TPE	Number of Elements	Percentage of Total
TPE 1A Subject-Specific Pedagogical Skills for Multiple-Subject Teaching Assignments	22	9.1
TPE 1B Subject-Specific Pedagogical Skills for Single-Subject Teaching Assignments	19	7.9
TPE 2 Monitoring Student Learning During Instruction	11	4.5
TPE 3 Interpretation and Use of Assessments	34	14.0
TPE 4 Making Content Accessible	24	9.9
TPE 5 Student Engagement	5	2.1
TPE 6A Developmentally Appropriate Practices in Grades K-3	6	2.5
TPE 6B Developmentally Appropriate Practices in Grades 4-8	11	4.5
TPE 6C Developmentally Appropriate Practices in Grades 9-12	4	1.7
TPE 7 Teaching English Learners	38	15.7
TPE 8 Learning About Students	15	6.2
TPE 9 Instructional Planning	31	12.8
TPE 10 Instructional Time	9	3.7
TPE 11 Social Environment	1	.4
TPE 13 Professional Growth	12	5.0
Total	242	100.0

To further support the content validity of the TE through the assessment development process, the program directors of the PACT IHEs were asked to judge the feasibility, task relevance, and representativeness of the TE for use within their institution. To examine the discipline-specific nature of the PACT assessments, the program directors vetted both the design and actual TE performances with their program faculty, soliciting their comments and critiques on the content validity of the Teaching Event and the reliability of the TE scores. All comments, concerns and issues were gathered and documented and, when appropriate, the PACT tasks were modified in response to this feedback.

Findings:

- There is a strong linkage between the TPE standards, the TE tasks and the skills and abilities that are needed for safe and competent professional practice.
- Each TPE except for TPE 12 (Professional, Ethical, and Legal Obligations) is represented in the Teaching Event. Like the developers of the California TPA, we determined that this TPE is more appropriately measured through other assessment methods.
- These findings provide strong evidence of the PACT Teaching Event’s content validity as a measure of beginning teacher competency.

The following section addresses the first part of the CCTC’s Assessment Quality Standard 19(f): *“The sponsor completes content review and editing procedures to ensure that pedagogical assessment tasks and directions to candidates are culturally and linguistically sensitive, fair and appropriate for candidates from diverse backgrounds. The sponsor ensures that groups of candidates interpret the pedagogical tasks and the assessment directions as intended by the designers...”*

Bias and Fairness Reviews

Bias/sensitivity review. The initial bias review followed guidelines put forth by Educational Testing Service (1998) for conducting bias/sensitivity reviews of assessments. The bias/sensitivity review process focused specifically on the Teaching Event handbooks and rubrics used in each certification area (i.e., multiple subjects, English/language arts, history/social studies, mathematics, and science).

18 experienced educators with expertise in detecting varied forms of bias were convened at Stanford University. After an orientation to the PACT assessment system and to the bias review guidelines put forth by ETS, panelists were given six questions to respond to that covered various forms of bias – stereotyping and language use, inflammatory or controversial material, inappropriate tone, contextual diversity, elitism, ethnocentrism, sexism, and inappropriate or irrelevant underlying assumptions. They first responded to each question individually, and then discussed their responses in a group. After the discussion, they answered the question again. The second set of ratings ranged from 70% to 100% of the participants indicating no bias related to any question in any assessment

task. Notes were taken on issues raised during the discussion, and the comments were noted along with the response during the revision process. See **Appendix C. PACT Bias and Sensitivity Review**, pp. 59-65 for tables summarizing reviewers' ratings on each task in the Teaching Event. Issues raised by reviewers were used to inform subsequent revisions of the 2003-04 Teaching Event handbooks, rubrics, and scorer training.

The following section addresses the CCTC's Assessment Quality Standard 19(g): *“The sponsor completes basic psychometric analyses to identify pedagogical assessment tasks and/or scoring scales that show differential effects in relation to candidates' race, ethnicity, language, gender or disability. When group pass-rate differences are found, the sponsor investigates to determine whether the differences are attributable to (a) inadequate representation of the TPEs in the pedagogical tasks and/or scoring scales, or (b) over-representation of irrelevant skills, knowledge or abilities in the tasks/scales. The sponsor acts promptly to maximize the fairness of the assessment for all groups of candidates and documents the analysis process, findings, and action taken.”*

Bias/fairness Analysis. In the studies described in the following section, the performance of candidates on the PACT assessment was examined to determine if candidates performed differentially with respect to specific demographic characteristics. Information was systematically collected to test for adverse impact with respect to the following indicators: gender, race/ethnicity, grade level, subject matter, demographic characteristics of the teacher's class (including proportion of English learners, economic backgrounds, and achievement levels) and school placement (context), and type of program (traditional preparation program vs. intern program). Where significant differential performances were found, these results were used to trigger further investigation to ensure that group differences truly reflect differences in the knowledge and skills being measured and were not solely a function of group membership.

To test for fairness across these demographic indicators, an ANOVA or t-test methodology was used. In the 2003-04 pilot, in addition to candidates' demographic backgrounds, we also studied other factors that might influence the assessment results such as the impact of community context of the teaching placement (e.g., urban, rural, suburban), reported level of family income in the candidate's class, and reported academic achievement in the candidate's class. In the 2003-04 pilot year, we had matched survey and score data for 287 candidates, representing 46% of the 625 scored Teaching Events.

In 2003-04 we found no significant differences between groups on the following factors: ethnicity of candidates⁷, reported percent of ELL students in candidates' classrooms, grade level taught (elementary versus secondary), and reported academic achievement level of candidates' students. In contrast to the previous year's results where no significant differences were found across groups, in the 2003-04 score data we found small differences in the total mean item scores between female and male candidates (with females scoring higher than males) and between candidates teaching in different

⁷ The groups large enough to evaluate included Latino/a, Asian, and white candidates. Sample sizes were not large enough for African Americans and Native Americans to allow comparisons.

community contexts (with candidates in suburban placements scoring slightly higher than those in urban and inner city placements). However, there were no statistically significant differences found in scores of candidates with different percentages of English learners in their classes. PACT will continue to monitor and re-examine the scorer training process and the design of the TE assessments to uncover any potential sources of bias that may exist due to varying socio-economic contexts or other demographic characteristics if they arise.

Given the disproportionate representation of females in our overall sample and the relatively small number of males, the gender differences could be due to the effect of outliers on the sample variance for males. One alternative explanation for the differences among candidates in different teaching contexts is that candidates teaching in urban settings were more likely to report (in the PACT Candidate Survey) the presence of 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 Teaching Event. 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.

Annual bias analyses will be conducted to continuously monitor possible sources of bias related to the Teaching Event.

Findings:

- Results of the bias and fairness review for 2002-03 handbooks and rubrics indicate that across the assessment tasks, a range of 70% to 100% of the reviewers found no bias related to any prompts in the tasks or in the scoring rubrics. See **Appendix C**, pp.59-65 for detailed results of this review.
- For the 2003-04 pilot, no significant differences were found in total mean item scores between candidates in different groups, including ethnicity of candidates, percent of English learners in candidates' classrooms, grade level taught (elementary versus secondary), and reported academic achievement level of candidates' students. However, there were marginally significant differences in scores between male and female candidates (with females scoring higher) and between candidates teaching in schools in different socio-economic contexts (with candidates in suburban schools scoring higher than those in urban or inner-city schools).

Annual bias analyses will be conducted to continuously monitor possible sources of bias related to the Teaching Event.

The following sections on Construct Validity and Concurrent Validity address the second part of the CCTC's Assessment Quality Standard 19 (b): *“Initially and periodically, the sponsor analyzes the assessment tasks and scoring scales to ensure that they yield important evidence that ... serves as a basis for determining entry-level pedagogical competence to teach the curriculum and student population of California’s K-12 public schools. The sponsor records the basis and results of each analysis, and modifies the tasks and scales as needed.”* **A description of the modification made to the tasks and scales is described in the section Assessment Design (page 11).**

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 can be defined as an idea developed to permit categorization and description of some directly observable behavior (Crocker & Algina, 1986).

Factor Analysis. The structure of the PACT scoring protocol was designed around the PIAR tasks in the TE to provide a framework for assessing teacher competence in relationship to the TPEs. Factor analysis was used to support the hypothesis that the PIAR teaching tasks (Planning, Instruction, Assessment, and Reflection) were meaningfully represented in relationship to candidate performance in the 2003-04 pilot. It was hypothesized that if the TE categories (PIAR) represent important dimensions of teaching, therefore, the structure of the data should cluster around these dimensions to support the category (PIAR) structure of the TE.

Because of the high number of Teaching Events scored in Elementary Literacy and Elementary Mathematics in the 2003-04 pilot (more than 100 each), these two areas were the focus of a factor analysis to determine whether item scores clustered into patterns, and if so, whether they fell into the PIAR categories. (Audit scores and calibration scores were excluded to eliminate duplicate Teaching Events.) Using the Principal Component Analysis extraction method and the Varimax rotation method, EL and EM scores (all rubric item scores, including double scores) were analyzed for factors.

Separate factor analyses of the Elementary Literacy (EL) and Elementary Mathematics scores resulted in two factors emerging, with the first factor being composed of Planning, Instruction, and Academic Language items, and the second factor being composed of Assessment, and Reflection items. These results were then replicated in independent analyses of the Secondary Mathematics, Secondary History-Social Science, and Secondary Science scores. Similarly, a factor analysis of all scores (which had common items across all subject areas) resulted in two factors similar to those obtained using subject-specific scores alone.

Table 6. EL - Rotated Component Matrix (2003-04 Pilot Year)

Guiding Question	Component	
	1	2
Planning GQ5. How well are the learning goals, instruction and assessments aligned?	.767	.324
Instruction GQ1. How does the candidate actively engage students in their own understanding of relevant skills and strategies to comprehend and/or compose text?	.761	.199
Planning GQ2. How does the instructional design reflect a coherent approach to the literacy curriculum?	.754	.341
Planning GQ1. How does the instructional design make the curriculum accessible to the students in the class?	.725	.231
Instruction GQ2. How does the candidate monitor student learning and respond to student comments, questions, and needs?	.710	.710
Planning GQ4. How does the instructional design reflect and address student interests and needs?	.701	.701
Planning GQ3. How does the instructional design reflect a balanced approach to the literacy curriculum?	.664	.392
Academic Language GQ1. How does the candidate's planning, instruction, and assessment support academic language development?	.548	.540
Reflection GQ1. To what extent did the candidate's reflections focus on student learning?	.314	.820
Reflection GQ2. What is the relationship between the candidate's reflections on teaching and on learning?	.161	.727
Assessment GQ2. How does the candidate analyze the two students' progress over time?	.298	.763
Assessment GQ1. How does the candidate's analysis of whole class learning reveal students' understanding of literacy?	.402	.722
Assessment GQ3. What is the quality of oral and written feedback to the two students about literacy?	.356	.662

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.
Notes: (a) Rotation converged in 3 iterations; (b) N=207

This section also addresses the CCTC's Assessment Quality Standard 20(h): *"To ensure that the overall construct being assessed is cohesive, the sponsor demonstrates that scores on each pedagogical task are sufficiently correlated with overall scores on the remaining tasks in the assessment."*

In addition to factor analysis, the correlation between mean task scores was also computed to examine how mean scores across tasks were related. Task scores are moderately to highly correlated. **Table 7** below displays the correlation between task scores for the 2003-04 pilot year. There was a moderate to high level of correlation of mean item scores across the tasks, suggesting that scorers can differentiate their judgments of teacher competence across scoring tasks. It also suggests that teachers who perform well on one task will also perform well at the others; i.e., scores on one of the Teaching Event Tasks will be highly predictive of scores on the other tasks.

**Table 7. Correlations between Teaching Event Task Mean Item Scores (PIAR)
2003-04 Pilot Year**

	Planning	Instruction	Assessment	Reflection	Academic Language
Planning	-				
Instruction	.707**	-			
Assessment	.706**	.654**	-		
Reflection	.646**	.596**	.753**	-	
Academic Language	.667**	.631**	.633**	.613**	-

** Correlation is significant at the 0.01 level (2-tailed).

Findings:

- In the 2003-04 pilot year, two factors emerged from the common rubric item scores, where factor one was comprised of Planning, Instruction, and Academic Language rubrics, and factor two was comprised of Assessment and Reflection rubrics.
- These results suggest that the assessment tasks (PIAR) are meaningful constructs that represent significant domains of teaching skills.
- There was a moderate to high level of correlation of mean item scores across the tasks, suggesting that scorers can differentiate their judgments of teacher competence across scoring tasks and that there is reasonable cohesiveness across the dimensions of teaching.

The following section addresses the CCTC’s Assessment Quality Standard 20(a): *“In relation to the major domains of the TPEs, the pedagogical assessment tasks and the associated directions to candidates are designed to yield enough evidence for an overall judgment of each candidate’s pedagogical qualifications for a Preliminary Teaching Credential. The program sponsor will document sufficiency of candidate performance evidence through thorough field-testing of pedagogical tasks, scoring scales, and directions to candidates.”*

Criterion-Related Concurrent Validity

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 performance on the Teaching Event can differentiate between candidates who are judged effective (meeting or exceeding performance standards) from those candidates who are judged less effective (not meeting performance standards). While this is one of the major purposes of the Teaching Event, studies of concurrent and/or criterion-related validity are seldom included in the validation of licensure tests (Poggio and Glassnapp et al., 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.

One study that PACT has conducted to explore the relationship of the PACT scores to other measures of candidates' competence was a comparison of scorers' analytic ratings with their holistic ratings of the candidate. Using the proposed cut score model resulting from the standard setting process described later in this document, we were able to compare candidates' pass/fail status (based on their Teaching Event scores and the passing standard proposed subsequently) with the holistic judgments made by scorers as to whether candidates should or should not be recommended for a credential based on reading their Teaching Events. The research hypothesis for this study was that the pass/fail decision based on candidate's scores on the Teaching Event and the cut score model should be consistent with the scorers' holistic ratings of the candidate's performance. While based on the same body of evidence, the passing standard was adopted long after the holistic ratings were made by scorers, so the TE's pass/fail status was independent of the scorer's holistic impression.

To gather data on the holistic judgment of the scorers regarding candidate performance, each scorer was asked to step back immediately after scoring a TE and holistically evaluate the candidate's performance based on the following question and rating scale:

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)"

To examine the relationship between the holistic rating and the candidate's Teaching Event score on the TE, each candidate's pass/fail status was computed using the cut score model proposed by the Standard Setting Panel and Confirmatory Group. The percentage of matches (Fail the TE and Holistic Rating=1, or Pass the TE and Holistic Rating>1) was computed to determine the level of agreement between these two sets of evaluative ratings.

Table 8 below displays the distribution of Holistic Ratings across subject areas for the 2003-04 pilot year. The percentage of candidates not recommended for a credential ranged across subject areas, from a low of 3.1% in Science to a high of 5.3% in Elementary Literacy.

Table 8. Holistic Rating by Subject Area (2003-04 Pilot Year)

Recommendation	SUBJECT AREA						TOTAL
	EL	EM	ELA	MTH	HSS	SCI	
Do not recommend	7 5.3%	1 1.2%	4 5.4%	1 3.0%	1 3.1%	1 2.4%	15 3.8%
Recommend	46 34.6%	44 51.2%	29 39.2%	16 48.5%	16 50.0%	15 35.7%	166 41.5%
Strong Recommendation –Solid	71 53.4%	35 40.7%	31 41.9%	15 45.5%	13 40.6%	23 54.8%	188 47.0%
Strong Recommendation - Exceptional	9 6.8%	6 7.0%	10 13.5%	1 3.0%	2 6.3%	3 7.1%	31 7.8%
Total	133 100.0%	86 100.0%	74 100.0%	33 100.0%	32 100.0%	42 100.0%	400 100.0%

Note: Only local campus scores were utilized in this analysis. 122 out of 522 Teaching Event scores were missing a Holistic Rating.

Table 9 below displays the 2003-04 pass/fail rates for Teaching Events across subject areas based on the passing standard proposed during Standard Setting in 2006-07. The passing standard that was accepted was the following:

Candidates pass the Teaching Event if they pass all five rubric categories (Planning, Instruction, Assessment, Reflection, and Academic Language) AND have no more than 3 failing scores of “1” across the tasks.

For the purpose of this analysis, the passing standard was modified to reflect the slight differences in the rubrics used in the 2003-04 pilot year and the rubrics used for the 2006-07 Standard Setting decision. For this analysis, Academic Language is not regarded as a task in itself because there is only one Academic Language rubric in the 2003-04 rubrics, and candidates must have no more than 4 “1”s rather than 3 “1”s.

Table 9. Pass/Fail Decision by Subject Area based on Final Adopted Passing Standard (adapted to account for differences in rubric structures) - 2003-04 Pilot Year

Met the Passing Standard ^a	SUBJECT AREA						TOTAL
	EL	EM	ELA	MTH	HSS	SCI	
Not Pass	25 13.4%	16 17.2%	18 18.0%	5 12.5%	7 15.9%	5 8.6%	76 14.6%
Pass	162 86.6%	77 82.8%	82 82.0%	35 87.5%	37 84.1%	53 91.4%	446 85.4%

Note: (a) The passing standard is based on the following rule – Pass if pass all four tasks (PIAR) AND Total #1s is not greater than 4. (To pass each task: Plan MIS \geq 1.6, Instruct MIS \geq 1.5, Assess MIS \geq 1.5, Reflect MIS \geq 1.5.) (b) All Teaching Event scores, including those that were double scores, calibration scores, and audit scores were included in this analysis. When there were multiple scores for a single TE, the average scores were used.

Table 10 below displays the decision consistency between the Holistic Rating and the Pass/Fail decision based on the approved passing standard. The sum of exact matches in the pass/fail decision (3.8% Not Recommended AND Fail + 85.0% Recommended AND

Pass) is 88.8%. This level of consistency (exact matches) is high and provides evidence of the validity of the assessment to judge candidate performance and to some degree the passing standard selected by Standard Setting participants (see Passing Standard on page 40). The percentage of mismatches for candidates recommended for a credential in the holistic rating but whose Teaching Event scores did not meet the passing standard (11.3%) requires further study. However, because these pass-fail data are based on a modified version of the passing standard, this level of discrepancy is unlikely to be as high in data from recent pilot years. Accuracy at the cut point is further addressed in our scoring procedures by double and triple scoring of scores near the cut-point.

Table 10. Crosstabs – Pass/Fail Decision and Holistic Rating

Holistic Rating	Pass/Fail Based on Passing Standard		
	Fail	Pass	Total
Not Recommended for Credential	15 (3.8%)	0 (0%)	15 (3.8%)
Recommended for Credential	45 (11.3%)	340 (85.0%)	385 (96.3%)
Total	60 (15.0%)	340 (85.0%)	400 (100.0%)

The following section addresses the first part of CCTC’s Assessment Quality Standard 20(h): *“The sponsor investigates and documents the consistency of scores among and across assessors and across successive administrations of the assessment, with particular focus on the reliability of scores at and near the adopted passing standard.”*

Consistency and Reliability

Traditionally, reliability is an examination of the consistency between a set of independent observations that are interchangeable. Psychometrically, reliability is defined as “the degree to which test scores are free from errors of measurement... measurement error reduces the reliability (and therefore the generalizability) of the scores obtained for a person from a single measurement” (AERA/APA/NCME, 1985, p.19). The complex structure of the PACT assessment system (the Teaching Event & Embedded Signature Assessments) poses some interesting challenges for conducting reliability analyses. In the PACT assessment, candidates respond to open-ended portfolio prompts about the design of their lessons and independently select work samples that are illustrative of their teaching. A candidate’s response to the portfolio tasks is complex, representing an integration of content knowledge and teaching skills (pedagogical content knowledge) within each subject area that shape the decisions the candidate makes about their instruction. Additionally, the teaching performance is not independent of the specific context of candidates’ classroom assignments and school placements.

The PACT consortium developed a scoring system that takes into account relevant contextual issues that could impact a candidate's performance. Therefore, achieving an acceptable level of consistency across scorers becomes a function of both the quality of the training of scorers and the moderation processes used to check scorer accuracy. Scorers are trained to follow the assessment procedures and to use the assessment tools (note taking forms and rubrics) consistently within and across TEs. To check rater consistency, a proportion (23%) of all locally scored TEs from the 2003-04 pilot were independently audited at a central Audit session.

Moderation processes in performance assessment are often based on examining the level of agreement between raters. Agreement statistics typically focus on the percent of paired ratings that are in perfect agreement, the percent that are in adjacent agreement, the percent that differ by 2 rubric points and the percent that differ by 3 rubric points. Additionally, moderation processes are also used to flag "borderline or problem TEs". That is, they are used to trigger a second (or third) independent read for the purpose of obtaining multiple perspectives on difficult-to-score TEs or on a TE performance that may lead to a recommendation against awarding a teaching credential.

To examine the reliability of PACT scores for the 2003-04 pilot, the following analyses were conducted. These analyses are consistent with those documented in the ETS TPA's score analysis report. To determine the extent to which local scores agreed with the audit scores, inter-rater agreement percentages were computed within each subject area. **Table 11** below displays the results of an analysis of agreement of audited Teaching Events across all subject areas.

When there was only one audit score, this was considered the "Audit Score". However, when there were three or more scores for a single Teaching Event (e.g., because of calibration scores at the local or audit scoring sites, double scores at the local or audit scoring sites, or an adjudicating score because of a two or more point discrepancy between the local and audit score), the "majority score" (the most frequent score that comprised more than 50% of all the scores for a given item) became the "Audit Score." If there was no "majority score" for a given item, it was considered a "missing score." The "Audit Scores" are the scores that are compared with the local campus scores in **Table 11** below.

Table 11 indicates that there was about a 91% level of exact agreement or agreement within one point. Overall, there were 12 adjudicated Teaching Events (8.2% of 146 total Teaching Events audited) at the Audit Score Session during this pilot year (whenever a Teaching Event was scored a "1" on **three or more rubrics** from a single scorer, it was double-scored).

When the ratings of local campus scorers are compared with the ratings of audit scorers on the same Teaching Events, we find that the ratings of local campus scorers was higher than those given by audit scorers by less than two-tenths of a point (.2). Although small, the direction of this difference was generally consistent and was statistically significant. This difference in mean scores was to be expected, given that data from a 2002-03

centralized scoring model compared to local scoring models suggested that there is moderate score inflation under local scoring models.

Table 11. Campus Scores – Audit Scores Consensus Estimate (2003-04 Pilot Year)

Items	Exact Match	±1	±2	±3	Missing**	Subtotals
Planning 1	102	67	17	2	15	203
Planning 2	111	53	15	4	20	203
Planning 3	107	73	14	1	8	203
Planning 4	106	64	15	3	15	203
Planning 5	107	67	14	2	13	203
Instruction 1	99	63	17	2	22	203
Instruction 2	110	66	17	3	7	203
Assessment 1	107	67	14	2	13	203
Assessment 2	102	60	21	3	17	203
Assessment 3	105	51	23	1	23	203
Reflection 1	107	75	11	3	7	203
Reflection 2	115	64	14	2	8	203
Academic Language	113	63	13	2	12	203
TOTALS	1391	833	205	30	180	2639
VALID PERCENT*	56.57%	33.88%	8.34%	1.22%		100.00%

Notes: (a) N=203 sets of Teaching Event scores. Every set of scores for Teaching Events that were scored more than once at the campus level (double-scored and calibration Teaching Events) was compared to the Audit Scores. (b) When there were multiple sets of scores for one Teaching Event (in either the Audit or Campus-assigned scores), all item scores for that Teaching Event (including Campus-assigned scores) were compared and the “majority score” for each item became the “Audit Score”. No Audit Score was assigned when there was no “majority score”.

*Percentage is valid percent (number of matches/total number of valid scores), excluding missing scores

** Most of the missing scores consist of score sets that represent a distance of 1 point (that could not be reconciled due to the lack of a “majority” score).

The following section addresses the last part of CCTC’s Assessment Quality Standard 19(f): “The sponsor ensures that... assessment results are consistently reliable for each major group of candidates.”

Reliability of scores across demographic groups. Reliability analyses across demographic groups using the 2003-04 field test data were limited by insufficient and unrepresentative sample sizes, due to low response rates on the demographic survey in that year. However, there were demographic data available from the previous pilot year (2002-03) that could be linked to double-scored Teaching Events. To satisfy Assessment Design Standard 19(f), we drew from the first year pilot data to conduct these analyses. There were 590 respondents on the demographic survey (about a 90% response rate) and 163 sets of score pairs. The merging of demographic survey responses and double-scored Teaching Events resulted in a sample of 149 cases. To test for bias in scoring reliability across demographic groups, consensus estimates were conducted for each of the major demographic groups by race/ethnicity (White, Asian, Hispanic/Latino⁸), primary language (English or Other), gender, and reported proportion

⁸ Insufficient sample sizes in the other ethnic groups did not permit consensus estimates for African Americans, Native Americans, and “Other”.

of ELL (English Language Learners) students in class (1=0%, 2=1-33%, 3=34-66%, 4=67-100%). The consensus estimates were similar for White and Hispanic/Latino candidates (about 90% of score pairs were exact matches or within one point), while 94% of score pairs were exact matches or within one point for Asian candidates. The consensus estimates for candidates who reported that their primary language was English or Other were also similar (90% of score pairs were exact matches or within one point for candidates who reported English as their primary language, while 92% of score pairs were exact matches or within one point for candidates who reported another language as their primary language). The consensus estimate for male candidates was slightly lower (88%) than for females (91%), but this difference is small. Last, the consensus estimates for candidates reporting varying proportions of English Language Learners in class did not vary and were at or very close to 90% of score pairs being exact matches or within one point.

Findings:

- In the 2003-04 pilot year, we found that about 91% of score pairs on subcomponents of the assessment were in exact agreement or agreement within one point.
- In the 2002-03 pilot year data, we found very little variation in the level of consensus agreement for different demographic groups of teachers, suggesting that there is no bias in the scoring reliability across groups.

Assessor Reliability

During the 2003-04 pilot year, 146 Teaching Events (23% of the 628 Teaching Events scored locally across content areas and campuses) were scored at the Audit Session (in some cases multiple times, if the Teaching Event was used as a calibration Teaching Event or if it was double scored). (See **Table 8**.) The reliability estimates below are based on double-scored Teaching Events in 2003 and includes only those Teaching Events scored both locally and at the Audit Session in 2004. **Table 12** below displays the assessor reliability for each task and overall assessor reliability for the 2003-04 pilot year, as well as the standard error of scoring (SES) for each task and for all the rubrics overall. The statistic used to estimate inter-rater reliability below is the Spearman Brown Prophecy reliability estimate (which was used by ETS to report on inter-rater reliability in the California TPA field review – See “Scoring Analysis for the Field Review of the California Teaching Performance Assessment, July 2003”). The formula described in the ETS scoring analysis is reproduced below:

$$\text{Assessor Reliability (task level)} = \frac{2 \times \text{corr}(A_1, A_2)}{1 + \text{corr}(A_1, A_2)}$$

where $\text{corr}(A_1, A_2)$ is the correlation between the first and second assessors’ scores. The reliability index ranges from zero to one, with increasing values indicating improving reliability. The standard error of scoring (SES) was calculated as:

$$SES = \sqrt{Var_i \times (1 - R_i)}$$

Where Var_i is the variance in the task Mean Item Scores for task i and R_i is the assessor reliability for task i . The SES quantifies the level of uncertainty associated with assessors' judgments (with reference to the consistency of scores assigned to candidates by different assessors scoring the same task. The smaller the SES, the less likely it is that the task scores received by teacher candidates are a function of the particular assessors who scored the tasks.

Table 12 below displays the assessor reliability for each task and the overall assessor reliability for the 2003-04 pilot year, as well as the standard error of scoring (SES) for each task and overall.

Table 12. Assessor Reliability and SES of the Teaching Event Tasks, 2003-04 Pilot Year

TASK-BASED RELIABILITY*			
Task	Correlation	Reliability	SES
PLANNING	0.490	0.658	0.411
INSTRUCTION	0.481	0.650	0.457
ASSESSMENT	0.595	0.746	0.416
REFLECTION	0.567	0.724	0.416
ACADEMIC LANGUAGE	0.500	0.667	0.481
OVERALL	0.545	0.880	0.977

*To estimate the task-based reliability, the Pearson-Brown formula was utilized to obtain the correlation between Task (PIAR-AL) Mean Item Scores.

Assessor reliability for the 2003-04 pilot year ranged from 0.650 in Instruction to 0.746 in Assessment. The overall assessor reliability was 0.880.

The following section addresses the CCTC's Assessment Quality Standard 19(i): *"In the course of developing or adopted a passing standard that is demonstrably equivalent to or more rigorous than the State recommended standard, the sponsor secures and reflects on the considered judgment of teachers, the supervisors of teachers, the support providers of new teachers, and other preparers of teachers regarding necessary and acceptable levels of proficiency on the part of entry-level teachers. The sponsor periodically re-considers the reasonableness of the scoring scales and established passing standard."*

PACT Standard Setting 2005-2007 Summary of Procedures and the Adopted Passing Standard

Standard setting is necessary to determine the consequential validity⁹ of the PACT Teaching Event, in other words, whether it validly distinguishes between candidates that

⁹ Messick, who developed the concept of consequential validity, defined the consequential aspect of validity as an appraisal of "the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (see Messick's essay on "Validity" in the 1989 (3rd) edition of Educational Measurement,

should or should not be recommended for a credential. Now that a passing standard for the Teaching Event has been established, studies that examine the relationship between passing/failing scores on the Teaching Event and other measures of teacher competency (e.g., supervisor ratings of performance, credentialing decisions within a program, certification assessments) can be conducted. The passing standard also allows us to call attention to Teaching Events that signal a need to go through a remediation process to help borderline candidates improve their scores, or to earmark borderline Teaching Events that need to be re-scored multiple times to ensure the reliability of the rubric scores.

The standard setting model that the PACT management team adopted was informed by standard setting models described by Dr. Edward Haertel, a member of the PACT Technical Advisory Group, in Haertel (2002) and Haertel & Lorie (2000), as described below. However, because the PACT Teaching Event is a performance assessment and is scored using rubrics (on a four point scale), the statistical procedures used in standard setting processes with traditional standardized assessments could not be used. Rather, we looked to the standard setting processes used by the National Board for Professional Teaching Standards (Phillips, 1996) and the Connecticut Beginning Educator Support and Training program, both of which score teacher performances on rubrics, for appropriate models of the standard-setting process.

We selected an evidence-based standard setting model that was informed by Haertel's three-stage model. In this three-stage process, we first convened a panel of teacher educators from the PACT consortium programs familiar with the Teaching Event rubrics and the scoring process to formulate initial recommendations for a passing standard on the Teaching Event. The second stage of the process called for a group of program leaders from the PACT consortium programs to review the same materials and the panel's recommendations and to select a set of passing standards that would be submitted to all participating programs in the consortium to review for acceptance. The last stage of the process involved convening program directors and deans from each of the consortium programs to accept or revise the proposed passing standard.

The **Final Passing Standard** that was ultimately selected by program directors and deans in January 2007 is below.

Candidates pass the Teaching Event if they pass ALL FIVE rubric categories (Planning, Instruction, Assessment, Reflection, and Academic Language) AND have no more than 3 scores of "1" across the tasks.

The cut score for each category is as follows: 1.66 for Planning (1 out of 3 scores can be a "1"); 1.5 in Instruction, Assessment, Reflection, and Academic Language (1 out of 2 scores can be a "1")

In other words, there are two ways a candidate could fail the Teaching Event

- 1) Fail one or more rubric categories
- 2) Have more than 3 failing rubric scores of “1” in categories across the Teaching Event

The following sections describe the standard setting procedures that began in 2005, and the final passing standard that was selected as a result of this process in January 2007. Consortium programs that do not accept the established passing standard may choose to establish their own passing standards. Programs that set their own passing standards will need to conduct their own research to validate those standards, subject to the approval of the CCTC.

Stage One: Standard Setting Panel. In early December of 2004, one member from each of the 16 PACT consortium campuses was invited to participate in the panel. The criteria used to select panelists included: 1) familiarity with the PACT handbook, rubrics, and the scoring process; 2) experience scoring Teaching Events; and 3) experience working with elementary credential candidates. This first stage of the standard setting process focused on the Elementary Literacy handbook and rubrics because of the large number of Teaching Events completed in that area in the 2003-04 pilot year. 2004 scores from the Elementary Literacy Teaching Events were utilized to simulate pass rates based on the recommended passing standard. Thus, familiarity or experience with scoring EL Teaching Events was also preferred in panelists. If invited panelists were unable to attend, their program directors were asked to send an alternate representative. The Standard Setting Panel Meeting was held on January 24, 2005 at Stanford University. Panelists from 12 of the consortium campuses participated in the process. Participants in the Panel Meeting are listed in **Appendix D: Timeline of PACT Activities & Participants**, pp.89.

To orient them to the standard setting process, participants in the panel were first introduced to the California TPA Passing Standard to distinguish between the rubric score levels utilized in the state’s TPA and the rubric score levels of the PACT rubrics. While the state’s TPA rubric level “1” is defined as representing “little or nothing to demonstrate competency”, the PACT level “1” is framed as having some strengths and weaknesses, but with a clear need for more work in order to meet the standard on each rubric. A rubric score of “2” on the PACT rubrics indicates a passing score on a rubric. Thus, in recommending a passing standard for the Teaching Event, panelists would have to keep in mind that a score of “1” indicates that a candidate has not met the minimum standard on a particular rubric item.

In addition, panelists were introduced to the Assessment Quality Standards (Program Standard 19 in the *Standards of Quality and Effectiveness for Professional Teacher Preparation Programs*). These documents were included in a “briefing book” that also included other resources, such as the Elementary Literacy Teaching Event handbook and rubrics, a TPE-Guiding Question Map, and other materials that panelists would use to formulate a passing standard.

After exploring their perspectives on what constitutes competent performance for a beginning teacher eligible to receive a state teaching credential, panelists reviewed nine

“candidate score profiles” that were compiled from the 2004 pilot year scores of Elementary Literacy Teaching Events. (A sample profile is displayed on page 43.) These score profiles were selected strategically from actual EL Teaching Event scores to represent performances that might lead to agreement on decision rules arising from discussions of each candidate’s performance on each of the five categories and on their overall performance. In addition to displaying the numerical rubric scores, profiles included the corresponding text from the rubrics.

Panelists were directed to focus on these descriptions of candidates’ performances rather than on the rubric scores alone in order to get a concrete idea of each candidate’s performance prior to making judgments about the performances. After reading each profile, panelists completed a Candidate Performance Review Form on which they evaluated each candidate’s performance on the five tasks and overall. The directions and rating descriptors for this task are below:

<i>Evaluate the candidate’s overall performance on the Teaching Event and on each of the PIAR tasks and explain the most important reasons for the evaluation based on the Guiding Questions, TPEs, or any aspect of the performance that most influenced your decisions.</i>	
<input type="checkbox"/> Below the passing standard	<input type="checkbox"/> Barely meets the passing standard
<input type="checkbox"/> Barely below the passing standard	<input type="checkbox"/> Surpasses the passing standard
Explain your rationale:	

Panelists were also prompted to explain the rationale for their evaluations of each task and the overall performance in order to capture the criteria they used in arriving at their evaluations of performance. As panelists completed their evaluations of each of the nine candidate profiles, their review forms were collected and their evaluations of candidates’ performances were tallied and displayed. A discussion on each of the candidate profiles followed, focusing on profiles in which there was some disagreement over whether a performance could be considered passing or not passing. In addition, when there was overall consensus about whether a performance was passing or failing, panelists were prompted to discuss their judgments and to explain their pass/fail judgments in order to articulate some decision rules that captured their evaluation criteria.

Sample Candidate Score Profile

P1	P2	P3	P4	P5	I1	I2	A1	A2	A3	R1	R2	AL
2	3	3	3	3	2	2	2	3	2	3	2	2

In terms of planning,

- P1 The candidate plans strategies to make the literacy curriculum accessible¹. These strategies have connections to the learners and/or the literacy curriculum content.
- P2 The instructional design includes a progression of activities and assessments that build understanding of the central concept(s), essential questions, or key skills that were identified as the literacy focus of the learning segment
- P3 The collection of instructional tasks or the set of assessment tasks make solid connections among facts, conventions, skills, and strategies in literacy.
- P4 The instructional design includes activities that draw upon aspects of students' backgrounds, interests, prior learning, or experiences to help students reach the literacy learning goals. Explicit consideration has been made for particular student needs that require differentiation or strategic teaching decisions.
- P5 The informal and formal assessments planned provide multiple opportunities for students to demonstrate progress toward the literacy learning goals. The assessments allow students to show some depth of understanding or skill with respect to the learning goals. The instruction planned provides students opportunities to learn what is assessed.

In terms of instruction,

- I1 The candidate structures classroom norms and activities to provide opportunities for students to use relevant skills and strategies to comprehend and/or compose text.
- I2 The candidate's responses to student comments and questions represent reasonable attempts to improve student abilities to use relevant skills and strategies to comprehend and/or compose text.

In terms of assessment,

- A1 The analysis identifies what most students understood or learned and what some or most did not. The analysis is supported through appropriate references to student work samples.
- A2 The analysis describes what students have done well, what they have done to some extent, and/or what they need to know with respect to the central concept(s), essential question, or key skills in literacy. It notes several changes over time.
- A3 The literacy feedback identifies what was done well and areas for improvement.

In terms of reflection,

- R1 Reflections focus on consideration of specific details of what students did or did not understand about literacy and the extent to which teaching practices were or were not able to facilitate student understanding of literacy content and skills. Reflections are consistent with the assessment results described in the previous task.
- R2 The candidate proposes and/or identifies changes in teaching practice based on reasonable assumptions about how student learning of literacy was tied to planning, instruction, or assessment decisions.

In terms of academic language support (AL),

- Learning goals, instructional activities, and assessments focus in a general way on at least one component of academic language.
- The candidate is aware of student needs with respect to academic language and uses scaffolds or supports to address these needs, but the implementation of these scaffolds or supports or the materials used are not well-suited to help students develop proficiency in academic language.
- The candidate's analysis of oral and written performance for academic language development identifies specific student strengths and needs.

Selecting a cut score model. Five cut scores models that operationalized potential decision rules into algorithms were presented. Each cut score model displayed the decision rules, passing and failing Teaching Event scores, and overall pass/fail rates for Elementary Literacy Teaching Events under each model. Panelists were then asked to select one of the models (or to propose their own models) and make suggestions for ways to revise it to better reflect the decision rules on which they had agreed. Panelists agreed on a cut score model that they felt best captured their decision rules and deliberations:

Initial Cut Score Recommendation:

Pass Teaching Event IF Pass All Tasks, OR Pass All but One Task AND Total Mean Item Score (Equal PIAR-L Weighting) ≥ 2.0

Cut Scores for each PIAR-L Task:

Plan MIS ≥ 1.6 ; Instruct MIS ≥ 1.5 ; Assess MIS ≥ 1.66 ; Reflect MIS ≥ 1.5)

This cut score model is both conjunctive and compensatory:

- Candidates pass the Teaching Event overall if they pass all of the PIAR tasks OR if candidates pass all but one task AND the Total Mean Item Score falls above a certain cut score (e.g., 2.0).
- In this model, all tasks, including the Academic Language rubric item, are weighted equally regardless of the number of rubric items in each task. (Thus, although the Academic Language rubric alone is not the deciding factor in the overall pass/fail decision, it is weighted more heavily than the other rubric items in other tasks.)
- The model is compensatory because it allows candidates to compensate for a lower performance on one task with higher rubric scores on other tasks (as long as those scores are high enough to bring their Total Mean Item Score to at least a 2.0 – or some other minimum cut score).

Stage Two: Confirmatory Group Meeting. A group representing program leadership (program directors and/or other faculty with leadership roles within their programs) in the PACT programs was convened to confirm or revise the passing standard recommended by the initial standard setting panel. The commission of this Confirmatory Group was to vet the recommended passing standard and decisions rules formulated by the Stage One Panel, and to evaluate, modify, and/or adapt the initial cut score recommendation. They would also tackle some of the issues that were not resolved by the first panel (e.g., how to include the Academic Language rubric in the overall passing standard, whether a candidate can pass overall when scoring less than a “2” on an entire task that covers certain TPEs exclusively.) All credential program directors from the 16 consortium institutions were invited to participate in the Standard Setting Confirmatory Group meeting, held one month following the Panel meeting. Directors that were unable to attend were asked to designate a representative from their institutions who would be empowered to vote on behalf of the directors. If an institution sent more than one representative, one delegate was designated to vote on the final passing standard. 14 representatives from 12 institutions participated in the Standard Setting Confirmatory Group meeting held on February 23, 2005 at San Jose State University. (Participants are

listed in **Appendix D: Timeline of PACT Activities & Participants**, pp.90-91.) There was overlap in the membership of the Panel and the Confirmatory Group in three cases.

One week before the meeting, Confirmatory Group members were sent the meeting minutes from the Standard Setting Panel meeting along with the recommendations of the panelists and directed to review these materials prior to arriving at the Confirmatory Group meeting.

When the Confirmatory Group members arrived at the meeting, they were provided with their own “briefing book”, a set of materials that included documents related to the state law, the handbook and rubrics for the Elementary Literacy Teaching Event, the TPE-Guiding Question Map, the Candidate Score Profiles, the summary of panelists’ evaluations of the Candidate Score Profiles, the EL Teaching Event scores from the 2004 pilot and several cut score model options that reflected the recommendations of the Standard Setting Panel. To orient them to the task, the Confirmatory Group was prompted to review the California TPA Passing Standard and the Assessment Quality Standards. The Confirmatory Group members were then briefed on the work of the Standard Setting Panel and were walked through the decision rules that were recommended by the panel, as well as the unresolved issues related to the Academic Language rubric and TPE coverage. The Confirmatory Group discussed these issues and agreed that none of the tasks should be privileged over others and that all tasks should be weighted equally.

The Confirmatory Group reviewed several cut score models that captured the recommendations of the Panel and variations of that model. They were also shown the pass/fail rates that would result across content areas if those cut score models were simulated on the scores in other content areas. After discussing the models at some length, the group voted on each of the five models. 10 out of 11 of the members of the group selected a model that was essentially identical to the one selected by the Standard Setting Panel.

Stage Three: Policy Group Meeting. The final stage in the standard setting process involved convening a group of program directors and deans from the PACT consortium campuses. What distinguished this group from the previous group is that all were program directors or other program leaders with decision-making power within their institutions. This final meeting was held on December 8, 2005 at UC-Davis and was attended by 18 program directors/faculty from 13 consortium programs (representatives from USC, Stanford, and Mills College were absent). The meeting was also attended by a representative from the UC Office of the President, a member of the PACT Leadership Team, a guest from UC-Merced and three members of the PACT Management Team. (Participants are listed in **Appendix D: Timeline of PACT Activities & Participants**, pp.93-94.)

Like the two previous committees, the policy group was oriented to the state law regarding the TPA, CTC’s Assessment Quality Standards, and the CA TPA passing standard. They were also provided with an overview of the standard setting process and

recommendations of the two previous committees. They examined the score profiles of two borderline candidates who would have passed under the passing standard proposed and confirmed by the first two committees. After a brief discussion of these profiles, participants were provided with an overview of the passing standard approach recommended by the previous committees (a task-based approach) and alternative approaches. The policy group came to a consensus to accept the task-based approach recommended by the previous committees. The group was then given time to examine and discuss various models for the task-based approach. **The participants voted 10 to 4** in favor of a cut score model (see inset box below) that was similar to the model recommended by the previous committees, but with the inclusion of Academic Language (which is now comprised of two guiding questions in the 2005-06 pilot year rubrics¹⁰) as an entire rubric category.

TASK-BASED APPROACH (Pass All or Most Tasks)

Passing Standard Recommended by Panels I & II and Approved by the Policy Group

Candidates pass the Teaching Event overall if they pass all PIAR tasks AND Academic Language, OR if they pass all but one task (which includes Academic Language as a task) AND the Total Mean Item Score (with Equal PIAR-A Weighting) ≥ 2.0

As was agreed upon by the first two panels, the Policy Group affirmed the decision rule that passing a task means that a candidate has passed with a score of “2” on the majority (or at least half) of the guiding questions within a task.

Revision of the Passing Standard. At the end of the 2005-06 academic year, the passing standard was revisited because of anomalies in the pilot data. For example, the set of scores below (on the 11 rubrics that comprise the 2006-07 scoring rubrics) suggests that simply passing all tasks with one score of “1” in each category should lead to a passing score on the Teaching Event, even though the candidate’s Teaching Event had five “1”s across the five categories.

➤ PASSES under approved passing standard: 221 12 12 12 12 Total MIS: 1.53

To address the problematic face validity of the passing standard described above, all participants in the Policy Group Meeting were sent a memorandum on September 15, 2006 in which an alternative passing standard was proposed. Participants were asked to indicate their preference for the previously approved standard or a proposed alternative. Participants voted by email and all responses were in favor of the proposed alternative passing standard. In meetings with PACT program directors on October 20 and December 1, 2006, additional discussions of the proposed alternative ensued. There was a consensus among program directors that the previously approved standard should be revised and that the passing standard be strengthened to prevent candidates who fail *any* of the five rubric categories (Planning, Instruction, Assessment, Reflection, and

¹⁰ The holistic Academic Language rubric was broken into two separate Guiding Questions to provide greater measurement clarity and to strengthen diagnostic information resulting from the scores.

Academic Language) from passing the Teaching Event. In addition, candidates may not have more than 3 scores of “1” across the 11 rubrics (in the most current versions of the rubrics).

The **Final Passing Standard** that was ultimately selected by program directors and deans in January 2007 is described below:

Candidates pass the Teaching Event if they pass ALL FIVE rubric categories (Planning, Instruction, Assessment, Reflection, and Academic Language) AND have no more than 3 scores of “1” across the tasks.

The cut score for each category is as follows: 1.66 for Planning (1 out of 3 scores can be a “1”); 1.5 in Instruction, Assessment, Reflection, and Academic Language (1 out of 2 scores can be a “1”)

In other words, there are two ways a candidate could fail the Teaching Event

- 3) Fail one or more rubric categories
- 4) Have more than 3 failing rubric scores of “1” in categories across the Teaching Event

The estimated pass rates across the six major content areas under this passing standard (adapted for the 2003-04 score data based on differences between the rubric structures) are found in **Table 9** (on page 34) above. The estimated overall pass rate under this passing standard was 85.4%.

Summary

In this report, we have described the score-related reliability and validity studies that have been undertaken to examine the suitability of the TE for awarding a preliminary credential to prospective teachers. To assess the validity of the TE we have conducted a number of analyses, examining:

- the construct validity of the assessment by comparing the structure of the guiding question items to the results of a factor analysis,
- the bias or fairness of the TE by examining differences in the scores of different demographic groups, and
- the criterion-related concurrent validity of the assessment by comparing raters' holistic ratings of candidates to pass/fail rates.

In addition, a number of studies were conducted or are in process to examine the reliability of the TE scores and to identify sources of error that might influence scorer ratings. A longitudinal study, funded by the Carnegie Foundation, is currently underway that will examine the predictive validity of the PACT scores in terms of teacher effectiveness in the first years of teaching (measured by student achievement).

The following section addresses the CCTC's Assessment Quality Standard 20(f): *"The sponsor carefully plans successive administrations of the assessment to ensure consistency in elements that contribute to the reliability of scores and the accurate determination of each candidate's passing status, including consistency in the difficulty of pedagogical assessment tasks, levels of teaching proficiency that are reflected in the multi-level scoring scales, and the overall level of performance required by the Commission's recommended passing standard on the assessment."*

PACT has piloted the Teaching Event and Rubrics for four successive years (2002-03, 2003-04, 2004-05, 2005-06). This is the fifth year of piloting. This technical report is based primarily on the 2003-04 pilot year results. In pilot years 1, 2, and 5, score reliability has or will be evaluated. The Teaching Event handbooks and rubrics have undergone minor revisions in each year in response to feedback from teacher candidates, teacher educators, trainers, and scorers. The final passing standard was recently established in January of 2007 after several rounds of deliberation and field tests. The passing standard, handbooks, and rubrics will be continuously reviewed to ensure that the difficulty of the tasks and levels of teaching proficiency reflected in the rubrics, and overall level of performance embodied in the passing standard are fair and consistent across successive administrations. After each year of implementation, feedback will be solicited from trainers, scorers, and those implementing the PACT in their programs. In addition, surveys of teacher candidates will continue to be administered to collect their perspectives on their experiences and perceptions of the assessment tasks.

Bibliography/References

- American Educational Research Association, American Psychological Association & National Council of Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council for Educational Measurement (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.
- California Commission on Teacher Credentialing (CTCC). (2001). Assessment Quality Standards. Updated July 21, 2004. Accessed originally from www.ctc.ca.gov/educator-standards/AssessmentQualityStds-CAT-E.doc (no longer available at this site)
Can now be downloaded from:
<http://www.ctc.ca.gov/commission/agendas/2006-11/2006-11-agenda.html> (See Action 7A, pages PSC 7A-14 through PSC 7A-17) (Updated February 9, 2007)
- California Commission on Teacher Credentialing. (2002). SB2042: Professional preparation programs – Teaching Performance Assessment. Updated Feb.1, 2002.
www.ctc.ca.gov/SB2042/TPA_FAQ.html. (no longer available at this site)
Can now be downloaded from:
<http://www.ctc.ca.gov/notices/coded/030005/030005.html> (Updated February 27, 2003)
- Crocker, L. and Algina J. (1986). *Introduction to a classical and modern test theory*. Fort Worth, TX: Holt, Rhinehart, & Winston.
- Cronbach, L. (1990). *Essentials of psychological testing* (5th edition). New York: Harper and Row.
- Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer (Ed.) *Test validity*. Hillsdale, NJ: Erlbaum.
- Educational Testing Service (2007). Testing accommodations for test takers with disabilities general information about testing accommodations.
Accessed on February 16, 2007 from <http://www.ets.org/disability>
- Equal Employment Opportunity Commission, 1978. *Uniform Guidelines on Employee Selection Procedures*. Washington, DC: EEOC.
- Haertel, E.H. (2002). Standard setting as a participatory process: Implications for validation of standards-based accountability programs. *Educational measurement issues and practice*, 21(1), 16-22.
- Haertel, E.H. and W.A. Lorie (2000). Validating standards-based test score interpretations. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, April 2000.

- Jones, P. and Johnson, A. (2004). An examination of categorical versus interpretive scoring rubrics. Presented at the Annual Meeting of the American Educational Research Association. San Diego, April 2004.
- Jones, P. (2005). An examination of Teacher Preparation Program Standard 19a Requirements and the Performance Assessment for California Teachers. Unpublished manuscript.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance based assessment: Expectations and validation criteria. *Educational Researcher*, 20(8), 5-21.
- Messick, S. (1989). Validity. R.L. Linn (Ed.) *Educational measurement* (pp.13-103). Washington, DC: National Council of Measurement in Education and The American Council on Measurement in Education.
- Messick, S. (1992). The interplay of evidence and consequences in the validation of performance assessment. Paper presented at the Annual Meeting of AERA/NCME. San Francisco.
- Moss, P. (1993). Can There Be Validity Without Reliability? *Educational Researcher*, 23(2): 5-12.
- Wilkerson Judy R. & Lang William S. (2003). Portfolio, the Pied Piper of Teacher Certification Assessments: Legal and Psychometric Issues. *Education Policy Analysis Archives*, 11(45).
- Phillips, G.W. (1996). Technical issues in large-scale performance assessment. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics. [Imprint]
- 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.
- United States Department of Education (2003). Higher Education-Guidance on Standards, Assessments, and Accountability: Assessment. Accessed on February 16, 2007 from http://www.ed.gov/policy/elsec/guid/standardsassessment/guidance_pg4.html
- WestEd (2000). Job analysis study. Cited in CCTC (1997). *California Standards for the Teaching Profession: A description of professional practice for California teachers*, p.37. Accessed on February 17, 2007 from http://ww1.psd.k12.ca.us/pro_development/BTSA/cfasst/California_TS.pdf.
- 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). Accessed on July 30, 2005 from <http://www.asu.edu/epaa/v11n45/>.