Examining the Extremes:

High and Low Performance on a Teaching Performance Assessment for Licensure

By Judith Haymore Sandholtz & Lauren M. Shea

In all types of performances, ranging from athletic competitions to theatrical events, even casual observers typically recognize the particularly stellar or poor performers. For trained observers, such as athletic scouts or theater critics, identifying the exceptional performers at both ends of the continuum tends to be the easiest part of the job. Similarly, in assessing the teaching practice of preservice teacher candidates, we expect that observers, particularly trained observers, will readily identify those who are exceptionally effective or ineffective. We anticipate that university supervisors and mentor teachers will agree on who demonstrates extraordinary performance for a preservice candidate and who needs additional preparation before taking on solo classroom teaching responsibilities. We assume that candidates who exhibit outstanding skills in student teaching will excel on a teaching performance assessment and that those who fail the assessment will be those who struggle in student teaching.

Given that both teaching performance assessments and university supervisors' observations include direct evaluation of teaching practice, we anticipate agreement in identifying high and low performers. Identifying weak candidates is particularly critical to ensuring that beginning teachers do not earn licenses until they are com-

Judith Haymore Sandholtz is a professor in the School of Education and Lauren M. Shea is director of education, outreach, and diversity with the Center for Chemical Innovation, Chemistry at the Space-Time Limit, in the Department of Chemistry, both at the University of California, Irvine, Irvine, California. judith.sandholtz@uci.edu & lshea@uci.edu

petent and ready to teach full time. Both university supervisors' observations and teaching performance assessments aim to evaluate the competency of preservice teacher candidates, and both approaches prompt concerns among teacher educators about their use for licensing decisions. Given the importance of summative judgments about teacher candidates, concerns about the reliability and validity of both approaches are paramount. Researchers find that summative judgments based on student teaching observations fail to differentiate among levels of effectiveness (Arends, 2006a). Similarly, concerns about the reliability and predictive validity of teaching performance assessments need to be resolved (Pecheone & Chung, 2006) before moving to widespread adoption. In addition, both approaches require substantial financial and human resources. In times of funding shortages, questions arise about the need to conduct both performance assessments and supervisor evaluations, particularly if both approaches reach the same conclusion about a candidate's readiness for licensure.

In an earlier study, we explored the extent to which university supervisors' perspectives about candidates' performance corresponded with outcomes from a summative performance assessment (Sandholtz & Shea, 2012). We specifically examined the relationship between supervisors' predictions and teacher candidates' performance on a summative assessment based on a capstone teaching event, part of the Performance Assessment for California Teachers (PACT). We opted to compare predictions and performance for three reasons. First, all of the supervisors were trained scorers of PACT. Because the training, calibrating, predicting, and scoring took place within a 2-week period, the supervisors were in a mind-set that aligned with the PACT ratings of effective teaching. Using the PACT scoring as a basis for determining readiness to teach was fitting for that time period and appropriate for making predictions of performance. Second, supervisors did not use a standard instrument during classroom observations, and they did not all complete observations during the same week. Consequently, using predictions and scores allowed us to make comparisons for a large number of candidates with a single instrument from the same point in time. Third, the process of making predictions did not significantly impose on the supervisors' workloads yet provided supervisors' judgments about candidates' readiness for licensure at that point in the year.

In contrast to expectations, we found that university supervisors' predictions of their candidates' performance did not closely match the PACT scores and that inaccurate predictions were split between over- and underpredictions (for complete findings, see Sandholtz & Shea, 2012). Our findings in that study, combined with suggestions from other researchers, prompted us to examine high and low performance through an in-depth follow-up analysis. In this follow-up study, we focus on four specific subsets of teacher candidates: not only the groups of high and low performers but also the groups of predicted-high and predicted-low performers, which were not examined in the earlier research. Analysis of the predicted-low performers (and within that group, the predicted-to-fail candidates) is important because that group includes candidates whom supervisors do not think are ready to be licensed yet pass the assessment. This follow-up study also expands the data sources and includes not only PACT score data but also information from student transcripts and student teaching. In addition, this study includes additional analyses that, for example, examine specific areas in the PACT to determine where differences occurred.

We address the following questions: (a) Do academic background factors correspond with high or low performance on the PACT? (b) In what specific areas on the PACT do high- and low-performing candidates excel and fail? (c) To what extent do university supervisors accurately predict high and low performance on the PACT? To what extent do candidates whom university supervisors predict will fail the PACT end up passing?

Assessing Teacher Competency for Teacher Certification

The central aim of teacher preparation programs is to prepare candidates to become effective, certified classroom teachers. The central aim of teacher certification systems is to affirm that teachers who receive licenses are qualified to enter the teaching profession. Teacher licensing systems are typically designed to ensure a basic level of teacher qualification (National Commission on Teaching and America's Future, 1996). However, because teacher licensing is a state responsibility, requirements for obtaining a teaching credential vary across states. In some states, applicants to teacher preparation programs must have a minimum grade point average and pass standardized tests focusing on basic skills before being admitted to a program. Upon program completion, candidates then must pass state-mandated tests that measure content knowledge and professional knowledge to receive an initial license to teach. In other states, testing occurs only at the end of the teacher preparation programs. The pass scores that candidates must achieve on licensing tests serve a screening function aimed at preventing incompetent teachers from entering the profession (Goldhaber & Hansen, 2010). The tests also provide a means to hold teacher preparation programs accountable for preparing competent beginning teachers and to allow states to compare candidates graduating from different programs (D'Agostino & Powers, 2009).

The main form of testing for teacher certification is paper-and-pencil exams consisting primarily of multiple-choice questions (D'Agostino & Powers, 2009). Using these types of tests for credentialing purposes has raised a range of concerns, including (a) the lack of direct classroom observation (Goldhaber & Hansen, 2010), (b) the constructs being measured (Berliner, 2005), (c) the elimination of qualified candidates who may perform poorly on paper-and-pencil exams (Goodman, Arbona, & Dominguez de Rameriz, 2008), (d) the limited relationship between the content of the licensure tests and teacher education programs (Sawchuk, 2012), and (e) the assessment of lower level subject matter knowledge that is not directly relevant to teaching (Mitchell & Barth, 1999). The overarching concern about paper-and-pencil licensing exams is that teachers' test scores do not predict teaching performance (Berliner, 2005; Darling-Hammond, Wise, & Klein, 1999). Researchers question the value of licensing exams in assessing teaching effectiveness, particularly the extent to which the tests are authentic and valid in identifying effective teaching (Darling-Hammond et al., 1999; Mitchell, Robinson, Plake, & Knowles, 2001; Wilson & Youngs, 2005). In a meta-analysis of 123 studies, D'Agostino and Powers (2009) reported that test scores were "at best modestly related to teaching competence" (p. 146) and concluded that performance in preparation programs was a significantly better predictor of teaching skills. Researchers have also reported that the limited information about teacher effectiveness gained from licensing exams varies across different populations of teachers (Goldhaber & Hansen, 2010 ; Goodman et al., 2008; Wakefield, 2003). Given the high pass rates, some researchers question the value of the tests in identifying candidates who are not ready to be licensed classroom teachers. Because candidates' average scores on state-required licensing tests tend to be higher than pass scores set by the states, researchers contend the tests should be only a minimum screen and used with other entry mechanisms (Sawchuk, 2012).

In an increasing number of states, concerns about licensing exams have prompted a move toward adopting teaching performance assessments. A key advantage of performance-based assessments is their use of evidence from teaching practice (Mitchell et al., 2001; Pecheone & Chung, 2006; Porter, Youngs, & Odden, 2001). Performance-based assessments may include, for example, lesson plans, curricular materials, teaching artifacts, student work samples, video clips of teaching, narrative reflections, or self-analysis. By using evidence that comes directly from actual teaching, performance assessments address the concern that licensing exams need to be connected to classroom teaching. Beyond assessing a candidate's knowledge and skills, the documents provide evidence about how the candidate is using these skills in specific teaching and learning contexts (Darling-Hammond & Snyder, 2000). The documents also provide insight into how teachers reflect on their practice and adapt their instructional strategies to be more effective. Compared to paper-and-pencil tests, performance-based assessments more closely reflect a conception of teaching that recognizes the complex, changing situations that teachers encounter (National Board for Professional Teaching Standards, 1999; Richardson & Placier, 2001).

In keeping with the name, teaching performance assessments are designed to engage candidates in tasks that stem directly from what they do in their classrooms and thereby to judge candidates' teaching performance. Rather than focusing on knowledge per se, the assessments aim to evaluate how a candidate applies this knowledge in the act of teaching. Performance assessments also are connected to professional teaching standards that reflect consensus about the components of effective teaching (Arends, 2006b). The teacher assessment systems developed by professional organizations such as the Interstate Teacher Assessment and Support

Consortium (InTASC) and the National Board for Professional Teaching Standards (NBPTS) include performance-based assessments that stem from established standards for the teaching profession. Despite the focus on teaching practice, concerns about the reliability and predictive validity of teaching performance assessments need to be resolved (Pecheone & Chung, 2006). Other concerns about performance assessments include competing demands, extensive requirements, effects on the curricula of teacher education programs, potential harm to relationships essential for learning, and the human and financial resources required (Arends, 2006b; Delandshere & Arens, 2001; Margolis & Doring, 2013; Snyder, 2009; Zeichner, 2003)

University supervisors also assess candidates' effectiveness as classroom teachers, but typically through formative evaluations. Although supervisors' observations provide a view into candidates' teaching performance, relying on them for summative judgments about candidates' competence raises concerns. Three of these concerns relate to issues of validity and reliability: training, specificity of observation forms, and frequency of observations (Arends, 2006a). The training that university supervisors receive may be inadequate to achieve interrater agreement. The observation forms may not be tailored for specific disciplines or grade levels, and classroom observations may not be conducted regularly. Supervisor observations also do not allow comparisons of candidates graduating from different programs.

In contrast to paper-and-pencil exams, teaching performance assessments and university supervisors' observations include direct evaluation of teaching practice. Consequently, we would expect both forms of assessment to reach similar conclusions about candidates' overall competence and readiness to teach. In particular, we would anticipate similar identification of preservice candidates who are not yet qualified to be credentialed teachers. This study explores those assumptions by examining the extremes of high and low performance.

Performance Assessment for California Teachers

The PACT is one of several teaching performance assessment models approved by the California Commission on Teacher Credentialing. Developed by a consortium of universities, the PACT assessment is modeled after the portfolio assessments of the Connecticut State Department of Education, the InTASC, and the NBPTS. The assessment includes artifacts from teaching and written commentaries in which candidates describe their teaching context, analyze their classroom work, and explain the rationale for their actions. The PACT assessments focus on candidates' use of subject-specific pedagogy to promote student learning.

The PACT program includes two key components: (a) a formative evaluation based on embedded signature assessments developed by local teacher education programs and (b) a summative assessment based on a capstone teaching event. The teaching event involves subject-specific assessments of a candidate's competency in five areas or categories: planning, instruction, assessment, reflection, and academic language. Candidates plan and teach an instructional unit, or part of a unit, that is videotaped. Using the video, student work samples, and related artifacts for documentation, candidates analyze their teaching and their students' learning. Following analytic prompts, candidates describe and justify their decisions by explaining their reasoning and providing evidence to support their conclusions. The prompts help candidates consider how student learning is developed through instruction and how analysis of student learning informs teaching decisions both during the act of teaching and upon reflection. The capstone teaching event is designed not only to measure but also to promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions; the teaching event also aims to stimulate teacher reflection on practice (Pecheone & Chung, 2006). The teaching events and the scoring rubrics align with California's teaching standards for preservice teachers. The content-specific rubrics are organized according to two or three guiding questions under the five categories identified earlier. Table 1 identifies the focus of the guiding questions within each category at the time of data collection. For each guiding question, the scoring rubric includes descriptions of performance for each of four levels or scores. According to the implementation handbook (PACT Consortium, 2009), Level 1, the lowest level, is defined as not meeting performance standards. These candidates have some skill but need additional student teaching before they will be ready to be in charge of a classroom. Level 2 is considered an acceptable level of performance on the standards. These candidates are judged to have adequate knowledge and skills, with the expectation that they will improve with more support and experience. Level 3 is defined as an advanced level of performance on the standards relative to most beginners. Candidates at this level are judged to have a solid foundation of knowledge and skills. Level 4 is considered an outstanding

Table 1
Focus of Guiding Questions in PACT Rubrics

Category	Focus of guiding questions
Planning	Q1: Establishing a balanced instructional focus
	Q2: Making content accessible
	Q3: Designing assessments
Instruction	Q4: Engaging students in learning
	Q5: Monitoring student learning during instruction
Assessment	Q6: Analyzing student work from an assessment
	Q7: Using assessment to inform teaching
Reflection	Q8: Monitoring student progress
	Q9: Reflecting on learning
Academic language	Q10: Understanding language demands
	Q11: Supporting academic language

Note. These were the foci of the questions at the time of data collection.

and rare level of performance for a beginning teacher and is reserved for stellar candidates. This level offers candidates a sense of what they should be aiming for as they continue to develop as teachers.

To pass the PACT teaching event, candidates must pass all five categories on the rubric (planning, instruction, assessment, reflection, and academic language) and have no more than two failing scores of 1 across categories. To pass a category, candidates must have passing scores of 2 or higher on at least half of the questions within each category. For example, because the instruction category includes two questions, at least one of the two scores must be a 2 or higher. The planning category includes three questions; therefore at least two of the three scores must be a 2 or higher. Teaching events that do not meet the established passing standard are double-scored. Candidates who fail the teaching event have one opportunity to resubmit. Candidates who fail more than one category or who have more than two scores of 1 across categories must complete a new teaching event. Candidates who fail only one category may resubmit the specific components for that category rather than the entire teaching event.

To prepare to assess the teaching events, scorers complete a 2-day training in which they learn how to apply the scoring rubrics. These sessions are conducted by lead trainers. Teacher education programs send an individual to be trained by PACT as a lead trainer, or institutions might collaborate to develop a number of lead trainers. The training emphasizes what is used as sources of evidence, how to match evidence to the rubric level descriptors, and the distinctions between the four levels. Scorers are instructed to assign a score based on a preponderance of evidence at a particular level. In addition to the rubric descriptions, the consortium developed a document that assists trainers and scorers in understanding the distinctions between levels. The document provides an expanded description for scoring levels for each guiding question and describes differences between adjacent score levels and the related evidence. Scorers must meet a calibration standard each year before they are allowed to score.

Methods

Sample

This study focuses on a subset of candidates from an earlier study of 337 candidates enrolled in a California public university's teacher education program over a 2-year period (Sandholtz & Shea, 2012). Our subset includes candidates whose performance or predicted performance on the PACT placed them at the high or low end of the continuum of the larger group of candidates. Before candidates' performance assessments were scored, university supervisors predicted each of their advisees' performance on the PACT. They predicted rankings of 1 to 4 on each of the 11 questions, which resulted in predicted total scores ranging from a possible 11 to 44. All of the supervisors were trained scorers, but they did not teach courses for student teachers and were not directly involved in preparing candidates for the performance assessment. The supervisors' role was to provide support and guidance for student teachers in their designated classrooms; they completed formative classroom evaluations but did not assign the student teaching grades. Consequently, their predictions stemmed from their classroom observations of candidates and their overall knowledge about scoring for the PACT but were not based on candidates' work in courses or drafts of their teaching events. After completing training for PACT scoring and passing calibration standards, university supervisors predicted scores for their advisees and then received their assigned assessments to score. Except in rare cases which were not included in the research, supervisors did not score the teaching events of their own advisees. The training, calibrating, and scoring took place within 2 weeks.

In this study, we specifically focus on four groups: high performers on the PACT, low performers on the PACT, predicted-high performers, and predicted-low performers. To identify the candidates in each group, we used cutoff scores of 37 for high performance and 20 for low performance (out of a possible 44). The cutoff scores of 37 and 20 fell at the end of the second standard deviation of the total scores for the 337 candidates and meant that candidates received a ranking on at least one question that was at the lowest or highest end of the rubric scale. The low performers received one or more rankings of 1, and the high performers received one or more rankings of 4. In the group of 337 candidates, we identified 22 high performers with a total score of 37 or higher, 21 low performers with a total score of 20 or less, 12 candidates whose supervisors predicted they would score 37 or higher, and 15 candidates whose supervisors predicted they would score 20 or less (see Table 2). Within the predicted-low group of 15 candidates, we identified 11 cases in which the supervisors predicted not only low performance but failure. Using the PACT guidelines for passing the teaching event, we identified those candidates who were predicted to fail by examining the number of failing scores of 1 on individual questions and categories. Some candidates' scores placed them

Table 2
Distribution of Candidates

	High performers[a]	Predicted-high performers[b]	Low performers[c]	Predicted-low performers[d]	Predicted to fail[e]
Multiple subject	8	7	14	10	7
Single subject	14	5	7	5	4
Math	4	1	3	1	1
Science	3	0	1	1	0
Social science	0	2	2	2	2
English/French	3	0	0	0	0
Art/music	3	2	1	1	1

[a] *n*=22. [b] *n*=12. [c] *n*=21. [d] *n*=15. [e] *n*=11

in both the high and predicted-high groups or in both the low and predicted-low groups. Two high-performing candidates were also predicted-high performers, and four low-performing candidates were also predicted-low performers.

Data Collection and Analysis

Data were drawn from candidates' records and included (a) demographic and student teaching placement information, (b) student transcripts, (c) predicted scores for the PACT teaching event, and (d) actual scores on the PACT teaching event. The records provided to researchers included assigned case numbers to protect individual identities.

To examine if academic background factors corresponded with high or low performance on the teaching assessment, we gathered data from the high- and low-performing candidates' transcripts about factors related to both their undergraduate education and their graduate credential program. As students in a postbaccalaureate teacher credential program, candidates entered the program holding a bachelor's degree in a specific discipline. Consequently, we included candidates' undergraduate university, undergraduate major, and undergraduate grade point average (GPA) as academic background factors. We also included two academic factors from the graduate credential program: grades in student teaching and grades in methods courses. Grades in student teaching offer a potential indicator of effectiveness in classroom teaching that is not based solely on supervisor evaluations. In this particular program, a candidate's grade for the student teaching component is based on a range of evidence, including submitted lesson plans, professional conduct, supervisor observations, mentor teacher evaluations, and other assignments. The program coordinators (elementary or secondary), rather than the supervisors, assign the grades for student teaching. We examined grades in methods courses because the curricula and assignments for those courses are the most closely connected to classroom teaching activities. Candidates preparing to teach in elementary schools complete multiple methods courses, including mathematics, science, language arts, social studies, reading, visual and performing arts, and physical education. Candidates preparing to teach in secondary schools complete a subject-specific methods course as well as a course about reading and writing in secondary schools.

We examined academic background data to look for patterns in connection with high and low performance on the performance assessment. We performed t-tests for independent samples to determine statistical differences between the mean grades of high and low performers in student teaching, methods courses, and undergraduate programs. We then computed correlations to determine the association between the grades (student teaching, methods courses, undergraduate GPA) and performance on the PACT.

To determine the specific areas of PACT in which candidates excelled and failed, we identified the number of Level 4 rankings for the high and predicted-high performers and the number of Level 1 rankings for the low and predicted-low performers on each of the 11 questions. We subsequently looked for patterns both within and across the subgroups. To investigate the extent to which university supervisors accurately predicted high and low performance, we compared predicted scores and actual scores on the PACT teaching event for four groups of candidates: high performers, low performers, predicted-high performers, and predicted-low performers. As described, the predictions and scores included a ranking from 1 to 4 on each of 11 guiding questions that are grouped within the five categories. The rankings are defined as follows: Level 1, not meeting performance standards; Level 2, acceptable level of performance; Level 3, advanced level of performance relative to most beginners; Level 4, outstanding and rare level of performance for a beginning teacher (PACT Consortium, 2009). The total possible score ranged from 11 to 44. To determine the association between supervisors' predictions and PACT scores for candidates, we conducted paired samples correlations for the total scores and the 11 questions. Correlation coefficients were adjusted for multiple tests using Bonferroni's correction, effectively making the alpha level .004. To determine percentages of supervisors who did not accurately predict candidates' performance, we used a frequency of distribution of difference and determined the difference between actual and predicted scores for each candidate's total score and each question.

Results

In the following sections, we present the results for each research question. We first present data about academic background factors and the correlation with candidates' PACT scores. We then report findings about performance on specific areas of the PACT for each subgroup, high performers, low performers, predicted-high performers, and predicted-low performers, and discuss the extent to which supervisors accurately predicted candidates' scores. We examine supervisors' predictions about which candidates would fail the assessment in the predicted-low performers section.

Academic Background Factors

In terms of candidates' academic backgrounds, we examined data about undergraduate majors, universities from which candidates received undergraduate degrees, and undergraduate GPAs. We also examined two factors from the graduate teacher credential program: grades in student teaching and grades in instructional methods courses. We found no clear trends related to undergraduate major or undergraduate university among the group of high- and low-performing candidates. As displayed in Figure 1, high and low performers completed undergraduate majors across fields. The highest number of both low performers (*n*=10) and high performers (*n*=6) majored in a social science field. An equal number of high performers (*n*=5) and low performers (*n*=5) majored in a science, technology, engineering, or mathematics field. As displayed in Figure 2, the majority of the candidates (77%) attended

Figure 1
Undergraduate degree majors for high and low performers.
Undergraduate degree data missing for 1 low performer.

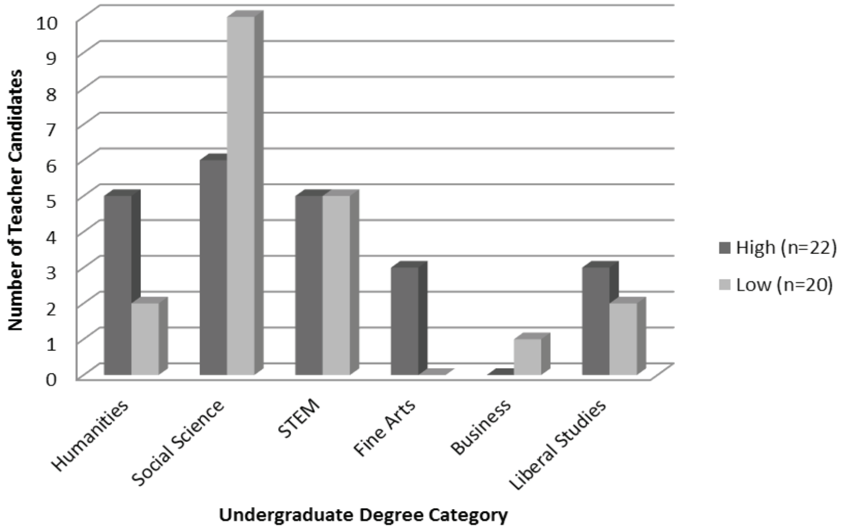
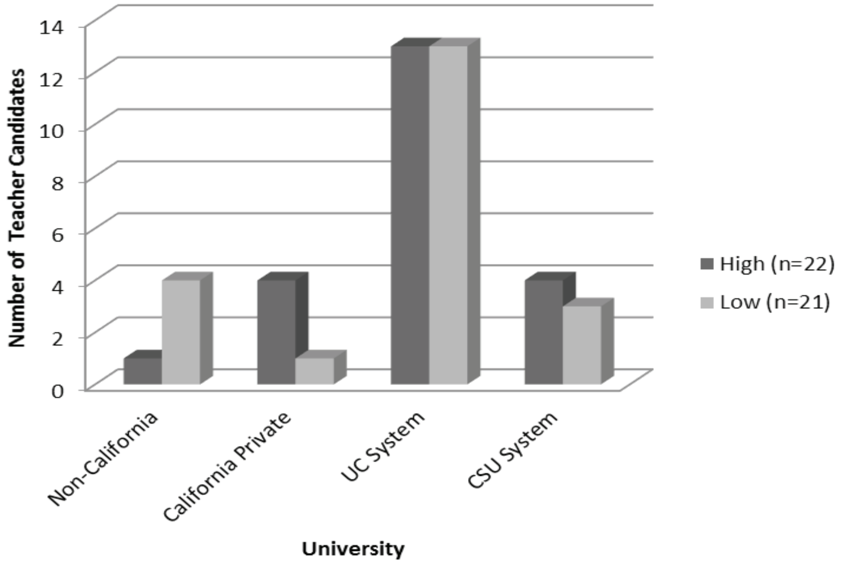


Figure 2
Undergraduate universities for high and low performers.

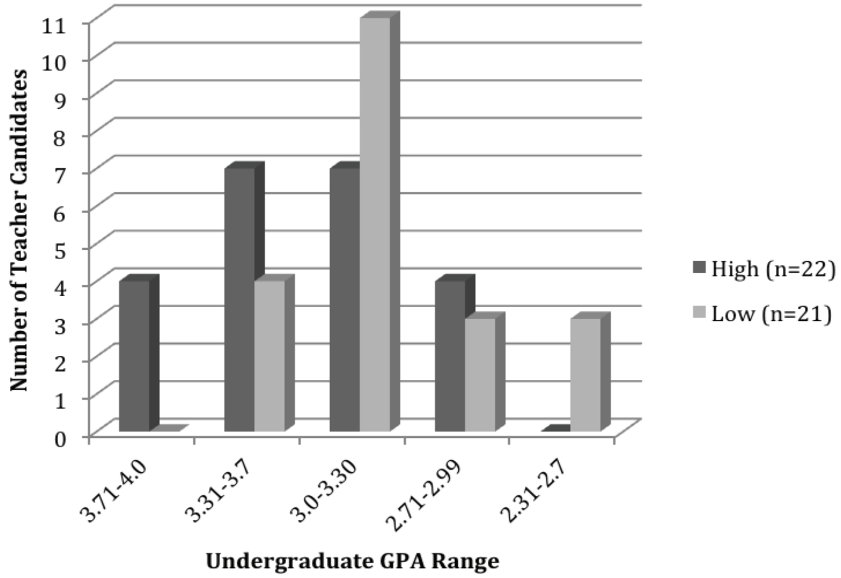


California public universities as undergraduates—an almost equivalent number of high performers (*n*=17) and low performers (*n*=16). Of the five candidates in the group who attended California private universities, four were high performers, and of the five candidates who attended non-California universities, four were low performers. Given the small numbers, we cannot suggest a trend in terms of private or non-California universities.

The undergraduate GPA for high performers (see Figure 3) ranged from 2.84 to 3.98 (*M*=3.36, *SD*=.36) and for low performers ranged from 2.45 to 3.66 (*M*=3.09, *SD*=.29). The mean for the high performers was significantly higher than the mean for the low performers. Twenty high performers and 13 low performers had student teaching grades of A (4.0); one low performer received a grade of D in student teaching (see Figure 4). Student teaching grades were nonsignificantly higher for high performers (*M*=3.96, *SD*=.094, range 3.70-4.0) than they were for low performers (*M*=3.68, *SD*=.77, range 0.70-4.0). High performers had significantly higher grades in their methods courses (*M*=3.97, *SD*=.054) than low performers (*M*=3.85, *SD*=.18).

Two academic background factors showed a moderate correlation with high and low performers' actual scores. As reported in Table 3, candidates' undergraduate GPAs and their grades in methods courses were significantly associated with performance on the PACT (.38 and .49, respectively). However, student teaching

Figure 3
Undergraduate GPAs of high and low performers.



grades showed no significant correlation with high-performing or low-performing candidates' actual scores on the PACT (*r*=.30).

Performance on the PACT

In the following sections, we report the specific areas in which candidates in each subgroup excelled and failed on the PACT. We also report the extent to which university supervisors accurately predicted candidates' high and low performance on the PACT and predicted those who would fail the assessment. Figure 5 displays the comparison of predictions and actual scores for the total sample of candidates in this study.

Figure 4
Student teaching grades for high and low performers.
Student teaching grade data missing for 1 candidate.

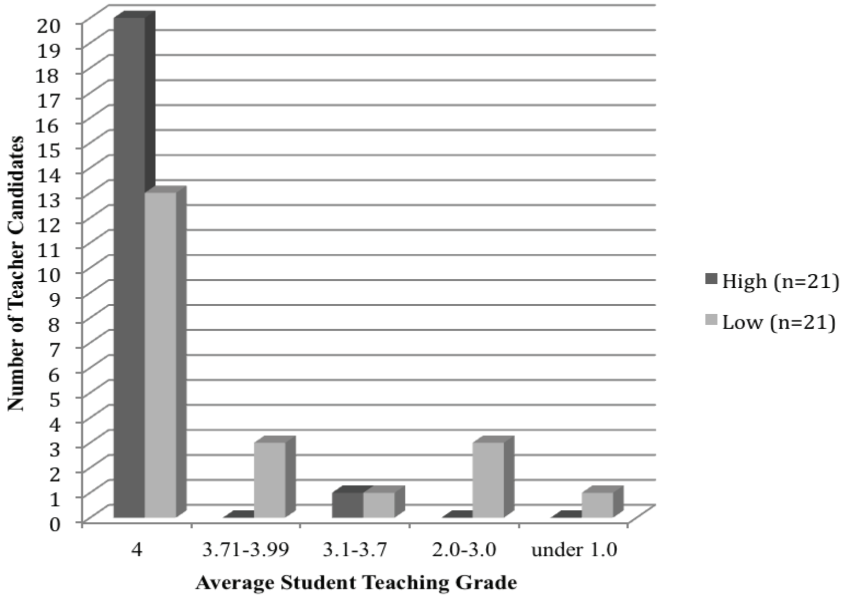


Table 3
Correlations of Actual PACT Scores to Teacher Candidate Academic Background Factors

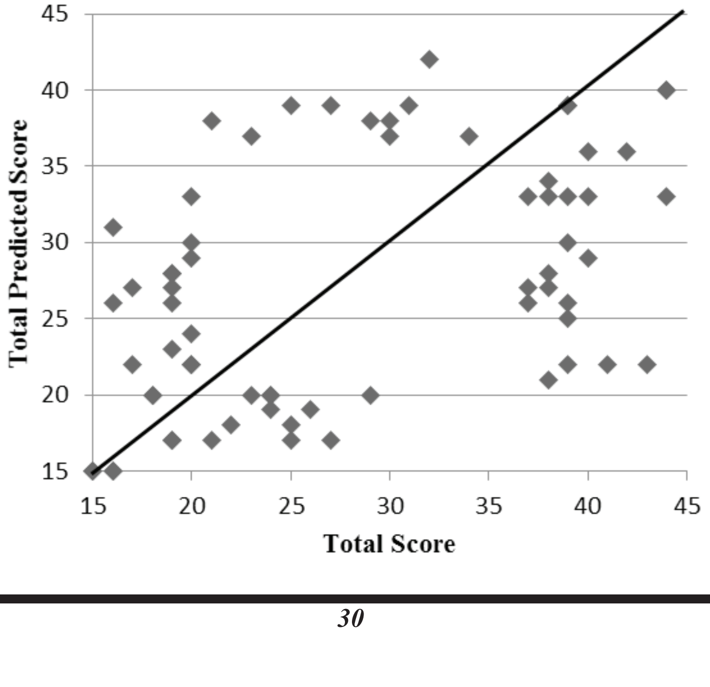
Teacher candidate academic background factor	Correlation to actual PACT score
Undergraduate GPA	.377*
Student teaching grade	.302
Methods courses grades	.486**

**p*<.05. ***p*<.01.

High performers. The specific areas in which the high performers scored at the highest level included two questions in the planning category and one in the assessment category. Eighteen of the 22 high performers (81.7%) received scores of 4 on Questions 1, 2, and 6. Question 1 focuses on how the plans for both learning tasks and assessment tasks support student learning. Question 2 examines if the plans make the curriculum accessible to the students in the class. Question 6 focuses on analyzing student work from an assessment and determines the extent to which candidates demonstrate an understanding of student performance with respect to standards or objectives. The questions on which the fewest high performers (~36%) received scores of 4 were Questions 10 and 11 in the academic language category. These questions examine if the candidate understands language demands and how the candidate's planning, instruction, and assessment support academic language development.

In the majority of cases, the supervisors did not predict the candidates' high performance on these questions. For example, on Question 6 (analyzing student work from an assessment), the supervisors underpredicted scores for 21 of the 22 high performers. Thirteen of these cases involved a 1-point underprediction, but in eight cases, the supervisors predicted the candidate would receive the lowest passing score, whereas the candidate received an exceptional score. On Questions 1 and 2 in the planning category, supervisors underpredicted performance for 64%

Figure 5
Comparison of predictions and actual scores for total sample.
Reference line represents matching scores or x=y.



and 77% of the high performers. Supervisors also underpredicted performance in the area of academic language, questions on which most of the high performers received scores of 3 or less.

In terms of total scores on the PACT, university supervisors were no more likely to accurately predict scores for high-performing candidates than for other candidates (Sandholtz & Shea, 2012). Total scores ranged from 37 to 44 for the high performers. The supervisors accurately predicted that all 22 high-performing candidates would pass the performance assessment. However, comparisons of total scores indicated that only one supervisor predicted an accurate total score for a high-performing candidate. For the other 21 high performers, university supervisors predicted total scores ranging from 20 to 40, underpredicting total scores by 4 to 21 points. In the two most extreme cases, supervisors underpredicted candidates' performance by nearly half of the possible total score, 19 points in one case and 21 points in the other. Twenty-two candidates received total scores of 37 or above; yet supervisors similarly predicted high performance in only two of those cases. As displayed in Table 4, we found no statistically significant correlations between predictions and total scores for high performers (r=.24). For individual questions, correlations ranged from −.364 to .404 in the high-performing group; none were statistically significant.

Low performers. The two areas in which the majority of the low performers (~70%) received failing scores of 1 were Questions 7 and 10. Question 7 focuses on assessment and how the candidate uses analysis of student learning to propose next steps in instruction. Question 10 examines how the candidate describes the language demands of the learning tasks and assessments in relation to student language development. In the majority of cases, supervisors did not predict the candidates' low performance on these questions. The supervisors predicted a passing score for 80% of those candidates who failed Question 7. In two of those cases, the supervisor predicted a score of 3, an advanced level of performance. On Question 10, supervisors predicted that 79% of those candidates who received failing scores of 1 would receive a passing score of 2.

Like the high performers, the low performers tended to score higher on Questions 1 and 2 in the planning category than on other questions. Only 3 of the 21 low-performing candidates received a failing score of 1 on Question 1 (establishing a balanced instructional focus), and only 5 of them received a failing score of 1 on Question 2 (making content accessible). But the supervisors predicted passing scores in each of these cases. For one low-performing candidate who failed both Questions 1 and 2, the supervisor had predicted an advanced score of 3. In another case, the supervisor predicted an exceptional score of 4 for all three questions in the planning category, but the candidate received failing scores of 1.

In terms of total scores on the PACT, we found no statistically significant correlations between predictions and total scores for low performers (r=.06). For

Table 4
Percentage of Accuracy and Correlations for Predictions and Scores for High Performers and Predicted-High Performers

	High performers[a] difference			Predicted-high performers[b] difference		
Question	0	±1	>1	0	±1	>1
Q1 Planning: Establishing a balanced instructional focus	36% (.279)	41%	22%	42% (.135)	50%	8%
Q2 Planning: Making content accessible	23% (−.364)	50%	27%	17% (.000)	66%	17%
Q3 Planning: Designing assessments	32% (.368)	50%	18%	42% (−.140)	25%	33%
Q4 Instruction: Engaging students in learning	41% (.212)	35%	23%	17% (.076)	58%	25%
Q5 Instruction: Monitoring student learning during instruction	27% (−.089)	40%	32%	33% (−.076)	41%	25%
Q6 Assessment: Analyzing student work from an assessment	5% (.404)	59%	36%	33% (−.451)	34%	33%%
Q7 Assessment: Using assessment to inform teaching	27% (−.162)	32%	41%	42% (.000)	33%	25%
Q8 Reflection: Monitoring student progress	32% (.058)	45%	23%	42% (.408)	50%	8%
Q9 Reflection: Reflecting on learning	50% (.176)	28%	22%	25% (.405)	42%	33%
Q10 Academic Language: Understanding language demands	32% (.240)	50%	18%	25% (.234)	66%	8%
Q11 Academic Language: Supporting academic language development	36% (.209)	50%	14%	42% (.319)	58%	0%
Total score	5% (.242)	0%	95%	8% (.371)	0%	92%

Note. Correlations for accurate predictions are in parentheses. Correlation tests were conducted using Bonferroni adjusted alpha levels of .004 per test (.05/12).

[a]*n*=22. [b]*n*=12.

individual questions, correlations ranged from −.311 to .718 in the low-performing group (see Table 5). The only statistically significant correlation (*r*=.72) was for Question 10 (understanding language demands). On this question, 5 of the 21 predictions for the low performers were accurate and the other 16 predictions were off by 1 point.

Predicted-high performers. In 12 cases, supervisors predicted that the candidates would do extremely well on the performance assessment (total scores of 37 or higher). For 75% or more of these predicted-high performers, supervisors anticipated exceptional scores of 4 on Questions 1, 2, 3, and 9. The first three questions are all in the planning category, suggesting that supervisors anticipated that these candidates would most likely excel in their planning for instruction. The questions focus on how the candidate's plans establish a balanced instructional focus, make the curriculum accessible to a variety of students, and include appropriately designed assessments. Question 9 asks how candidates use research, theory, and reflections on teaching and learning to guide their teaching practice. But only 15%-42% of the predicted-high performers received exceptional scores in these areas. On Questions 3 and 9, supervisors overpredicted scores by 2 points in one-third of the cases.

The area in which the supervisors did not anticipate exceptional performance for these candidates was Question 11, which focuses on how candidates' planning, instruction, and assessment support students' academic language development. Only 1 of the 12 predicted-high performers had a predicted score of 4 in this area; this candidate's actual score was 3. All of the predictions for Question 11 for predicted-high performers were within 1 point of the actual score, most frequently a predicted score of 3 and an actual score of 2.

The majority of candidates whom supervisors anticipated would score particularly high on the assessment did not receive total scores in the high performance range. As displayed in Table 4, there were no statistically significant correlations between predictions and total scores for predicted-high performers (*r*=.371). Only 2 of the 12 predicted-high performers actually received total scores in the high-performing range. In the remaining 10 cases, supervisors overpredicted total scores by 3 to 17 points for total score predictions ranging from 37 to 42.

All of the predicted-high performers passed the overall assessment, but two candidates received a score of 1 (not meeting the performance standard) on one question. In one case, the supervisor predicted a score of 3 (an advanced level of performance), and in the other case, the supervisor predicted a score of 4 (an outstanding and rare level of performance). For 10 of the 12 predicted-high performers, supervisors overpredicted the candidates' scores by 2 points on one or more questions. In one case, the supervisor overpredicted the candidate's score by 3 points on one question and 2 points on four other questions. On 42% of the questions for which supervisors predicted a score of 4 for these 10 candidates, the

Table 5

Percentage of Accuracy and Correlations for Predictions and Scores for Low Performers and Predicted-Low Performers

	Low performers[a] difference			Predicted-low performers[b] difference		
Question	0	±1	>1	0	±1	>1
Q1 Planning: Establishing a balanced instructional focus	43% (−.238)	43%	14%	40% (.202)	60%	0%
Q2 Planning: Making content accessible	33% (−.104)	57%	10%	60% (.031)	33%	7%
Q3 Planning: Designing assessments	43% (−.274)	38%	9%	40% (.286)	53%	7%
Q4 Instruction: Engaging students in learning	43% (.229)	52%	5%	27% (.142)	67%	7%
Q5 Instruction: Monitoring student learning during instruction	48% (.085)	48%	5%	53% (.026)	47%	0%
Q6 Assessment: Analyzing student work from an assessment	28% (−.311)	43%	28%	60% (.120)	33%	7%
Q7 Assessment: Using assessment to inform teaching	28% (.171)	62%	10%	60% (.342)	33%	7%
Q8 Reflection: Monitoring student progress	33% (.067)	52%	10%	60% (−.375)	33%	7%
Q9 Reflection: Reflecting on learning	52% (.494)	38%	10%	60% (.518)	40%	0%
Q10 Academic Language: Understanding language demands	24% (.718**)	76%	0%	67% (.592)	33%	0%
Q11 Academic Language: Supporting academic language development	52% (.077)	43%	5%	53% (.366)	47%	0%
Total score	5% (.059)	5%	90%	13% (.560)	0	87%

Note. Correlations for accurate predictions are in parentheses. Correlation tests were conducted using Bonferroni adjusted alpha levels of .004 per test (.05/12).

[a]*n*=21. [b]*n*=15.
**p<.004.

candidates received a score of 2. These 2-point and 3-point ranges mean that, in particular areas where supervisors predicted candidates would excel, they either failed to meet the performance standard or received the lowest passing score. For individual questions, the correlations ranged from −.451 to .408, and none were statistically significant in the group of predicted-high performers (see Table 4).

Predicted-low performers. Supervisors predicted that 15 candidates would receive total scores of 20 or lower (out of a possible 44 points), a score that is considered low performance in this study. A total score of 20 meant that the candidate received a 1, the lowest score on the rubric, on at least one question. For more than half of these predicted-low performers, supervisors anticipated failing scores of 1 on Questions 3, 4, 7, 9, and 10. These questions focus on designing assessments, engaging students in learning, using assessment to inform teaching, reflecting on learning, and understanding language demands. In contrast, none of the supervisors predicted these candidates would fail Question 1, in the planning category.

Predicted total scores for these 15 candidates ranged from 15 to 20, and actual total scores ranged from 15 to 27. The candidates whom supervisors anticipated would perform poorly on the assessment did not achieve high scores, but surprisingly, the majority did not fall in the low performance range. Of the 15 predicted-low performers, 4 (26.6%) actually received total scores of 20 or less. The remaining 11 candidates received total scores ranging from 21 to 27 points. Supervisors underpredicted these candidates' total scores by a range of 3 to 10 points.

Of the four subgroups in this study, the supervisors' predictions of total scores were closest to the actual scores for the predicted-low performers. However, as displayed in Table 5, there were no statistically significant correlations between predictions and total scores for predicted-low performers (r=.560). In addition, for individual questions, none of the correlations were statistically significant and ranged from −.375 to .592 in the group of predicted-low performers.

In 11 of the 15 cases of predicted-low performance, supervisors anticipated not only low performance but failure. That is, supervisors predicted that candidates would receive scores of 1 (not meeting performance standards) on three or more questions, which would constitute failing the performance assessment. However, only 3 of the 11 predicted-to-fail candidates (27.2%) actually received failing scores. The other eight candidates received no more than two scores of 1 on individual questions and passed the assessment with total scores ranging from 21 to 27 points. For most of those who were predicted to fail, the difference between the supervisors' predictions and their actual scores on any individual question was 1, with the supervisor predicting a failing score of 1 and the candidate receiving a passing score of 2. However, for four candidates, the supervisor predicted a failing score of 1 on an individual question, but the candidate received a score of 3. These 2-point underpredictions mean that the supervisor predicted the candidate would not meet the standard in that particular area but the candidate received a score indicating an advanced level of performance. In one case, the supervisor predicted a passing score of 2 on an individual question, but the candidate received the highest score of 4, considered an outstanding and rare level of performance.

For these 11 candidates, supervisors predicted failing scores in areas that span all five categories. For example, for 7 of 11 candidates (64%), supervisors predicted a failing score of 1 on Question 3 in the planning category, Question 4 in the instruction category, Question 7 in the assessment category, Question 9 in the reflection category, and Questions 10 and 11 in the academic language category. However, only 9%-36% of the predicted-to-fail candidates received failing scores on these questions. In contrast, supervisors did not predict failure for any of the predicted-to-fail candidates on Question 1, which focuses on the extent to which candidates' plans establish a balanced instructional focus. Only one of these candidates actually received a failing score of 1 on Question 1.

Discussion

Our first research question asked whether academic background factors correspond with high or low performance on the PACT. Our findings reveal a correlation between the high- and low-performing candidates' grades in university course work and their scores on the performance assessment; this correlation may reflect the academic elements of the PACT. Although the assessment focuses on classroom teaching, the format requires significant amounts of written analysis. Students who receive high grades in university courses likely possess strong literacy skills and analytical abilities. These same skills likely help teacher candidates in analyzing their teaching, communicating their reasoning in a written form, and providing evidence for their claims. The association we found for high and low performers between grades in methods courses in the credential program and scores on the PACT may indicate a similarity between course assignments and elements of the performance assessment. Instructors of methods courses often evaluate assignments in which candidates develop lesson plans, select instructional strategies, and provide the rationale for their instructional decisions. Similar to the performance assessment, these assignments may take the form of written documents, include some videotaped teaching segments, and involve critical analysis of the videotaped segments. Consequently, grades in methods courses may reflect students' abilities to accomplish the types of tasks included in both the methods courses and the PACT. The lack of correlation between student teaching grades and PACT scores in this study may manifest because the majority of high and low performers received grades of A in student teaching. In this program, student teaching grades are assigned by the program coordinator, not the supervisors, and are based on lesson plans, observation reports, mentor teacher evaluations, professional conduct, and other assignments. Consequently, student teaching grades may not necessarily correspond with a supervisor's perspectives about a candidate's effectiveness in the classroom.

Whereas supervisors' predictions of PACT scores varied across candidates, student teaching grades tended to be high.

Our second research question focused on identifying the specific areas in which high- and low-performing candidates excelled and failed on the PACT. Across subgroups, there appears to be a pattern of stronger performance, as well as higher predicted performance, on Questions 1 and 2 in the planning category. The high performers had the most rankings of 4 and the low performers had the fewest rankings of 1 on these two questions. In addition, for the predicted-low and the predicted-to-fail candidates, none of the supervisors predicted failure on Question 1. For the predicted-high performers, supervisors predicted the most rankings of 4 on questions in the planning category. These findings may reflect the importance of planning in effective teaching and the fact that candidates typically gain experience in instructional planning beginning early in a credential program. It would be highly unusual for a candidate to excel in the instruction category on the PACT but not the planning category. In contrast, candidates may develop appropriate plans but fail at enacting those plans in an active classroom. The category in which high-performing candidates had the fewest rankings of 4 and low-performing candidates had the most failing scores of 1 was academic language. Whereas candidates typically enter credential programs recognizing the need to plan for instruction, they may be unfamiliar with the role of academic language in student learning. Moreover, candidates must understand the language demands embedded in instructional activities before they can effectively support students in developing and using academic language.

Our third research question focused on the extent to which university supervisors accurately predict candidates' high and low performance on the PACT and accurately predict who will fail the assessment. Examining supervisors' predictions of their candidates' scores on the assessment provides a means of making direct comparisons with actual performance as well as a means of capturing supervisors' perspectives about candidates' readiness for licensure. Supervisors know performance on the PACT determines whether candidates will qualify for a teaching credential. When they predict that candidates will do particularly well on the summative assessment, supervisors are suggesting that candidates are highly qualified to assume full-time teaching responsibilities as credentialed teachers. When they predict that candidates will fail the summative assessment, they are indicating that candidates are not yet ready, in their view, to assume solo classroom teaching responsibilities. Because supervisors' predictions are not communicated to candidates and hold no weight in outcomes of the assessment, we think their predictions serve as a forthright measure of their perspectives about candidates' qualifications for licensure.

As reported in our earlier study (Sandholtz & Shea, 2012), we anticipated that supervisors who observe and assess candidates' classroom teaching would be well positioned to predict how individual candidates would perform on a teaching performance assessment and, in particular, would accurately predict which candidates would perform particularly well or poorly. However, in this study, whether we looked at candidates predicted to be high or low performers or candidates who actually were high or low performers, we found differences between supervisor predictions and actual scores on the performance assessment. In the group of 43 high- and low-performing candidates, supervisors predicted high or low performance in only 6 cases. Similarly, in the group of 27 predicted-high and predicted-low performers, only 6 candidates actually received scores in the high or low performance ranges. Moreover, the majority of candidates whose supervisors predicted failure did not fail, and the majority of candidates who did fail had been predicted to pass. We also found a surprising lack of agreement between predicted and actual scores on specific questions on which candidates excelled or were predicted to excel. Because supervisors review candidates' lesson plans in connection with classroom observations, one might anticipate that supervisors would make closer predictions on questions related to planning; however, that was not the case.

This apparent lack of agreement about candidates at both ends of the continuum is puzzling. As trained scorers who pass the PACT calibration standard each year, the supervisors are clearly knowledgeable about the assessment. Differences would not stem from predictions being made by people who do not understand the PACT. In addition, we found no evidence to support the theory that some scorers or supervisors may tend to be "easier graders" than others. When we examined cases in which the supervisor–scorer pairs were the same, we found differing ranges between predictions and scores. For example, in the cases of two low performers with the same supervisor and scorer, the prediction and score matched in one case but differed by 10 points in the other case. Similarly, in the cases of two high performers with the same supervisor and scorer, the prediction and score matched in one case but differed by 17 points in the other case. If a supervisor were consistently predicting higher scores, the range between predictions and scores for the same supervisor–scorer pairs would be similar across cases. When we examined the cases of high- and low-performing candidates with the largest differences between predictions and scores, we found that they had different supervisors, which also suggests that the score differences are not due to tendencies of a particular scorer or supervisor.

Differences in their assigned tasks may explain why scorers and supervisors do not always identify the same candidates as high and low performers. Although scorers and supervisors engage in the same general task of assessing candidates' teaching, they draw on different data sources, observe candidates in different contexts, and make judgments over different time frames (Sandholtz & Shea, 2012). They also may differ in their perspectives about high and low performers because of the extent of writing involved in the PACT. Because supervisors in this program do not teach seminars for student teachers or directly prepare candidates for the performance assessment, they typically do not encounter written assignments from their candidates, particularly not written analyses of their teaching. As part of classroom observations, supervisors have discussions with candidates about their plans and classroom instruction. Some candidates may be effective classroom teachers and adept at reflecting on their instructional practice in discussions with supervisors but not as skilled in writing about their planning and teaching.

Our findings suggest that identifying teacher candidates who are particularly effective or ineffective as classroom teachers is not as straightforward as we anticipated. University supervisors, methods course instructors, and scorers of performance assessments all may have differing perspectives about the competency of individual candidates. Candidates who exhibit outstanding skills in student teaching may not be those who excel on a performance assessment, and candidates who fail the assessment may not be those who struggle in student teaching.

The limitations of this study highlight potential areas for future research. Given the small sample size of high and low performers, statistical significance was difficult to reach. In addition, the study was limited to data from one teacher education program, which may have specific features that contributed to the findings. Research that includes multiple universities would yield a larger sample size and allow comparisons across teacher education programs. Research that follows candidates into the first years of teaching could examine the extent to which high and low scores on a performance assessment, or supervisors' predictions of performance, are associated with effective classroom teaching.

Conclusion and Implications

The findings of this study highlight four issues related to the assessment of preservice teacher candidates. First, our findings suggest that student teaching grades may not serve as discriminating forms of evaluation, even for candidates who perform particularly well or poorly on a teaching performance assessment. In line with other research reporting that the majority of candidates receive a grade of A in student teaching (Arends, 2006a), we found that the majority of both high and low performers in our study received a grade of A in student teaching. In many programs, grades in student teaching are assigned by the university supervisor and may be based largely on supervisors' evaluations of candidates; but in the program we studied, a coordinator assigned grades, and the supervisors' observations composed only a portion of the overall grade. In either type of arrangement, student teaching grades offer little information about candidates' qualifications if there is insufficient differentiation. In addition, a single letter grade provides no information about specific areas of strength and weakness.

Second, the results of this study prompt questions about the connection between candidates' academic strengths and classroom teaching performance. The association we found between candidates' grades in methods courses and their scores on the PACT, combined with the lack of association between candidates' predicted and actual scores, suggests that the academic requirements of the assessment may be as important as the teaching segments. A key advantage of performance assessments is the use of evidence that comes directly from actual teaching. However, because the format of the PACT involves written documents in which candidates provide analysis and explanations of their actions, candidates benefit from strong literacy skills in completing the tasks. Grades in methods courses reflect candidates' written work but not their enactment of plans in the classroom. It is unclear whether candidates may be effective teachers but not do as well on the performance assessment because they have less skill in writing about their planning and teaching. In future studies, researchers may want to examine the extent to which performance assessments emphasize candidates' academic abilities.

Third, the lack of agreement in identifying exceptional candidates at both ends of the continuum warrants further investigation. A key aim of teaching performance assessments is to identify candidates who are not adequately qualified and prepared to be licensed teachers. Although a majority of candidates may pass summative teaching performance assessments and earn teaching credentials, we need to be confident that an assessment is accurately identifying weak candidates who are not ready for solo classroom teaching. When candidates whom university supervisors predict will fail a summative performance assessment end up passing, we wonder what concerns about candidates' qualifications are not being identified in the assessment. Conversely, when candidates whom supervisors predict will pass the assessment end up failing, we wonder what weaknesses the assessment is capturing that the supervisors are not identifying. The differences in predictions and actual scores related to high performance are equally puzzling but do not hold the same implications for licensing decisions. Achieving an outstanding score on a performance assessment may be a point of personal pride for a candidate, but a high score does not influence the type of credential awarded or future job prospects.

Finally, the findings of our earlier study and this follow-up study raise questions about relying on a single measure to evaluate teacher candidates for licensing decisions. Different types of assessments may provide contrasting information about candidates' strengths and weaknesses. Candidates who fail a performance assessment may demonstrate competence in courses and supervisor observations of student teaching, and candidates who pass the assessment may demonstrate less effectiveness in courses and supervisors' observations. Both teaching performance assessments and university supervisor observations focus on direct assessment of teaching practice yet may reach different conclusions about a candidate's skills and progress. If there is variation across sources about which candidates are not yet qualified to receive a teaching credential, we may want to be cautious about making licensing decisions based on the outcome of a single measure. Given the complexity of teaching, assessment systems that include multiple sources of evidence may offer a more comprehensive appraisal of candidates' overall readiness to teach. Researchers studying teacher effectiveness conclude that no single factor can predict success in teaching (Peterson, 1987, 2000; Rockoff, Jacob, Kane, & Staiger, 2011). Different measures address different aspects of teacher quality, and multiple evaluators, who hold different roles, contribute varying perspectives about teacher quality (Peterson, 2000). Berliner (2005) contended that a single performance is inadequate for evaluating teacher quality. Using multiple measures to make summative judgments about teacher candidates seems prudent given the importance of licensing decisions and the possibility that different measures may identify different candidates as lacking the necessary qualifications to be credentialed teachers.

References

Arends, R. I. (2006a). Summative performance assessments. In S. Castle & B. S. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 93-123). Lanham, MD: Rowman & Littlefield Education.

Arends, R. I. (2006b). Performance assessment in perspective: History, opportunities, and challenges. In S. Castle & B. S. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 3-22). Lanham, MD: Rowman & Littlefield Education.

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*, 205-213. doi:10.1177/0022487105275904

D'Agostino, J. V., & Powers, S. J. (2009). Predicting teacher performance with test scores and grade point average: A meta-analysis. *American Educational Research Journal, 46*(1), 46-182. doi:10.3102/0002831208323280

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*, 523-545. doi:10.1016/S0742-051X(00)00015-9

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

Delandshere, G., & Arens, S. A. (2001). Representations of teaching and standards-based reform: Are we closing the debate about teacher education? *Teaching and Teacher Education, 17*, 547-566. doi:10.1016/S0742-051X(01)00013-0

Goldhaber, D., & Hansen, M. (2010). Race, gender and teacher testing: How informative a tool is teacher licensure testing? *American Educational Research Journal, 47*(1), 218-251. doi: 10.3102/0002831209348970

Goodman, G., Arbona, C., & Dominguez de Rameriz, R. (2008). High-stakes, minimum-competency exams: How competent are they for evaluating teacher competence? *Journal of Teacher Education, 59*(1), 24-39. doi: 10.1177/0022487107309972

Margolis, J., & Doring, A. (2013). National assessments for student teachers: Documenting teaching readiness to the tipping point. *Action in Teacher Education, 35*, 272-285. doi:10.1080/01626620.2013.827602

Mitchell, R., & Barth, P. (1999). *Not good enough: A content analysis of teacher licensing examinations*. Washington, DC: Education Trust

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. Washington, DC: National Academy Press.

National Board for Professional Teaching Standards. (1999). *What teachers should know and be able to do*. Arlington, VA: Author.

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

PACT Consortium. (2009). *Implementation handbook*. Retrieved from http://www.pacttpa.

org/_main/hub.php?pageName=Implementation_Handbook

Pecheone, R. L., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers (PACT). *Journal of Teacher Education, 57*(1), 22-36. doi:10.1177/0022487105284045

Peterson, K. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*, 311-317. doi:10.3102/00028312024002311

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Porter, A., Youngs, P., & Odden, A. (2001). Advances in teacher assessments and their use. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 259-297). Washington, DC: American Educational Research Association.

Richardson, V., & Placier, P. (2001). Teacher change. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 905-947). Washington, DC: American Educational Research Association.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43-74.

Sandholtz, J. H., & Shea, L. (2012). Predicting performance: A comparison of university supervisors' predictions and teacher candidates' scores on a teaching performance assessment. *Journal of Teacher Education, 63*(1), 39-50. doi:10.1177/0022487111421175

Sawchuk, S. (2012). Teachers pass license tests at high rates. *Education Week, 31*(19), 1-2.

Snyder, J. (2009). Taking stock of performance assessments in teaching. *Issues in Teacher Education, 19*(1), 7-11.

Wakefield, D. (2003). Screening teacher candidates: Problems with high-stakes testing. *The Educational Forum, 67*, 380-388. doi:10.1080/00131720308984587

Wilson, S., & Youngs, P. (2005). Research on accountability processes in teacher education. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying teacher education: The report of the AERA Panel on Research and Teacher Education* (pp. 591-643). Mahwah, NJ: Lawrence Erlbaum Associates.

Zeichner, K. M. (2003). The adequacies and inadequacies of three current strategies to recruit, prepare, and retain the best teachers for all students. *Teachers College Record, 105*, 490-519.

.