Theme Article

Examining the Internal Structure Evidence for the Performance Assessment for California Teachers: A Validation Study of the Elementary Literacy Teaching Event for Tier I Teacher Licensure

Journal of Teacher Education
2014, Vol. 65(5) 402–420

DOI: 10.1177/0022487114542517
jte.sagepub.com
SAGE

Brent Duckor[1], Katherine E. Castellano[2], Kip Téllez[3], Diah Wihardini[2], and Mark Wilson[2]

Abstract

Interpretations for licensure tests involve a series of inferences or a validity argument, leading from the test score to decisions about who is accepted or denied entry into a profession. Utilizing an argument-based framework for validation based on the Standards for Educational and Psychological Testing, we explore the evidence for the ongoing use of the Performance Assessment for California Teachers (PACT) for Tier I licensure decisions. The evidence for a unidimensional and a multidimensional structure based on the instrument's content are examined with an item response model. Examining operational data (n = 1,711) from seven California teacher education institutions, we found sufficient internal structure validity evidence to support the continued, but limited, use of this instrument for its intended summative purpose. Evidence for a three-dimensional structure of model fit better explains overall teacher candidate performance on the PACT instrument as it is currently designed.

Keywords

teacher licensure, internal structure evidence, item response modeling, multidimensionality, 1999 testing standards

Kane (1994) noted that each intended interpretation assigned to test scores needs to be supported with evidence. He writes that the interpretations for licensure and certification tests involve "sequences of inferences, or an argument, leading from the test score to decisions about licensure or certification" (p. 133). The argument-based framework for validation of test results has gained momentum in the last several decades (Cronbach, 1988; Kane, 2013; Haertel & Herman, 2005; Messick, 1989, 1994) as has the concern for consequences and the use of high stakes test score results for individuals and institutions (Haertel, 2013; Moss, 2013; Shepard, 1993). The renewed focus on the meaning of validity is both epistemological and pragmatic: Put sharply, what are the grounds for believing that a particular teacher licensure and certification score has meaning for the profession? Which particular lines of evidence are necessary (and which are merely sufficient) to warrant claims about preservice and inservice teachers' "skills," "proficiencies," "dispositions," and so forth? What is to be done with teacher licensure results when different types of validity evidence apparently conflict with one another? Most vexing to teacher educators, what happens when our validation efforts pose problems for deeply held convictions about the virtues of authentic assessment and portfolios in particular, and the promise of multifaceted task-based performance assessments more generally?[1]

These are the questions that motivated our study. If one views validation as we do, as a principled, pragmatic, and scientific activity that focuses on claims made by test developers, then teacher educators are obligated to know about the psychometric qualities of the assessments they demand for licensure. We point to documents such as the *Standards for Educational and Psychological Testing* (the *Testing*

[1]San Jose State University, CA, USA
[2]University of California, Berkeley, USA
[3]University of California, Santa Cruz, USA

Corresponding Author:
Brent Duckor, Associate Professor, Department of Secondary Education, Lurie College of Education, San Jose State University, SH 436, One Washington Square, San Jose, CA 95192-0074, USA.
Email: Brent.Duckor@sjsu.edu

Standards; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999), which suggest that we have a professional responsibility to engage with and monitor the validity evidence for any large-scale testing and examination system. In our view, the development and use of any teacher licensure instrumentation must be guided by the *Testing Standards* to ensure the "quality of the evidence"; moreover, "a few lines of solid evidence regarding a particular proposition are better than numerous lines of evidence of questionable quality" (p. 11).

In this article, we examine the meaningfulness of score results from the Performance Assessment for California Teachers (PACT) by investigating evidence for the hypothesized internal structure of the assessment, and, by extension, test content. According to the *Testing Standards*, evidence based on internal structure concerns the "degree to which the relationships among test items and test components conform to the construct on the proposed test score interpretations are based" (p. 13). The PACT instrument developers claim that teacher candidates who obtain a passing score possess the "skills and abilities" that are needed for safe and competent professional practice according to the standards of the teaching profession (Pecheone & Chung, 2007). According to the *Testing Standards*, evidence based on test content concerns documentation of the "themes, wording, and format of the items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring" (p. 11). Thus far, validation efforts have focused on the content of the PACT and not the internal structure of score results at scale.

The conceptual framework for any teacher licensure examination system may imply a single dimension of behavior or it may posit several components that are expected to be homogeneous, but that are also distinct from each other. The structure of the PACT, from a content validity perspective, suggests both possibilities. The teacher candidate obtains a single composite score to warrant her "skills and abilities," that is, to support the claim that she is ready for "safe and competent professional practice" in the California classroom (Pecheone & Chung, 2007).[2] The teacher candidate also receives a set of subscores for different tasks in each domain of practice (Planning, Instructing, Assessing, Reflecting, and Academic Language). Thus, the report invites interpretations of both the global score and the subscores.

Using a multidimensional analysis of PACT production (i.e., nonpilot) data, we investigate the extent to which these subscores are, in fact, interrelated. Moreover, we examine the empirical evidence for increasing task difficulty across the scales put forward by the PACT developers. Our emphasis on obtaining evidence for internal structure of the PACT is not a trivial one. The concern over multidimensionality in testing is driven, in part, by a concern for misinterpretation of results. If subsets of PACT tasks have a specific characteristic in common (e.g., language load, reading level, technological and writing demands), these may function differently for different groups of similarly scoring examinees. More worrisome, the reliability of subscores based on only a few tasks is notoriously low. To the extent policy makers, evaluators, and teacher educators *misinterpret* the meaning and generalizability of scores derived from large-scale instruments such as the PACT, the potential for unintended consequences multiply.

The Promise of Performance Assessment and Rationale for Summative Licensure Instruments for California Teachers

Performance assessments for credentialing preservice teachers, like the PACT Teaching Events (TEs), became mandatory in California with the passing of the 1998 California Senate Bill (SB) 2042. By law, all California teacher candidates must meet the California Standards for the Teaching Profession, or more specifically, the Teacher Performance Expectations, regardless of their type of accredited professional preparation. Accordingly, the California Commission on Teacher Credentialing (CCTC, 2012) and the Educational Testing Service developed a state-standardized assessment—the California Teacher Performance Assessment. SB 2042 also allows for alternative assessments to the state assessment as long as they are aligned with the California Standards. Thus, teacher preparation programs at several California higher education institutions, including universities from both the California State University and University of California systems, formed the PACT consortium to develop an alternative curriculum-embedded assessment using an "evidence-based system" (Pecheone & Chung, 2007).

Teacher performance assessments (e.g., PACT) are not the only requirements for licensure in California. There are other indicators to evaluate teacher candidate "skills," "abilities," and "proficiencies." Evidence from academic coursework, subject matter competency, and field observations are also used to make judgments about candidate proficiency with the California Standards. Nonetheless, the scores derived from these data sources do not necessarily meet technical criteria for validity, reliability, or generalizability in so-called evidence-based approaches to evaluation (Cochran-Smith, 2005).

Compounding the technical challenges facing the valid and reliable interpretation of "local" assessment data, different stakeholders appear to hold different views of what constitutes proficiency with the California Standards for the Teaching Profession. There is disagreement about the purposes and appropriate uses of the teacher candidate data itself. Wise and Leibbrand (2001) maintained that

> sound assessment systems are integrated with learning experiences throughout teacher candidates' development and are not merely a series of off-the-shelf measures. They are embedded in the preparation programs and conducted on a

> continuing basis. Candidate monitoring is planned in response to faculty decisions about the points in the program best suited to gathering performance information. (p. 3)

Despite the tensions between university faculty, school-based cooperating teachers, field placement supervisors, and teacher education program administrators around data interpretation and use, the need for summative, not merely formative, evaluation of student teachers is clear (Farkas & Johnson, 1997; Murray, 2001). Raths and Lyman (2003) found that far too many student teachers receive low-level formative evaluations throughout their program and yet earn a teaching degree and license because these formative evaluations fail to coalesce into a negative summative appraisal. Presumably the PACT and other similar summative performance assessments in California should prevent such possibilities and thus serve to uphold common standards in the teaching profession across the state.

The Validation Study: Research Questions, Data, Methods, and Analysis

Educational researchers and practitioners who work with teacher licensure exams can tackle the question "What is being tested in the PACT?" in various ways. We do so by framing the question in the larger context of a validation study based on the *Testing Standards* (AERA, APA, & NCME, 1999). Our approach uses psychometric modeling to investigate the internal structure of the PACT data for the Elementary Literacy Teaching Event. The study examines the meaningfulness and reliability of the PACT scores in two principal ways. First, our approach provides information on the distribution of teacher candidates' "skills and abilities" in a large cross-sectional sample. Second, the approach describes the location of items on the same scale by their "task difficulty." We also explore the relations between the content domains (as represented by the scoring rubrics) to determine if and how they are related to one another. Specifically, this study addresses the following two research questions consistent with the validation studies proposed by the *Testing Standards* (AERA, APA, & NCME, 1999):

> **Research Question 1:** To what extent does a scaling model fit the Elementary Literacy PACT data and help us better understand the relationship between teacher candidate "skills and abilities" and the difficulty of the items/tasks?

In particular, we address this question by determining the extent to which

a. the scaling model fits the data,
b. the results are consistent with expectations about the scale, and
c. the test scores are reliable.

> **Research Question 2:** Is there evidence that the Elementary Literacy PACT assesses the multiple "dimensions" of teaching as its test structure and the PACT technical report suggests?

In particular, we address this question by determining the extent that the following multidimensional models fit well enough to describe teacher candidate performance:

a. The unidimensional scaling model,
b. The task-based scaling model (determined by the structure of the tasks embedded in the PACT TE),
c. The domain-based scaling model (determined by the five domains purportedly assessed by the PACT TE), and
d. Other domain-based scaling models (such as planning/instructing/metacognition) driven by hypotheses now emerging from large-scale implementation.

Before addressing these research questions, we first provide more details about the PACT instrument. We then review previous validity and reliability studies to contextualize our current study and the need for more rigorous methods to uncover the problem of dimensionality embedded in the teacher candidate performances.

The PACT Instrument

For teaching certification in particular subject areas in California, the PACT instrument consists of a summative TE and formative Embedded Signature Assessments (ESA). The ESAs are specifically designed by each teacher education program and thus are not the focus of this more general study. In contrast, TEs are mandatory and consistent across all programs. They consist of five tasks evaluated across five domains—Planning, Instruction, Assessment, Reflection, and Academic Language—with two to three scored questions/items per domain for a total of 12 items. Figure 1 illustrates this structure, and Table A1 in the appendix provides the specific questions that correspond to each of the 12 domain scores for the Elementary Literacy Teaching Event in particular. According to the PACT technical report (Pecheone & Chung, 2007), the Planning, Instruction, Assessment, and Reflection domains reflect four stages of teaching, and their scores are drawn from materials unique to tasks related to each stage. As shown in Figure 1, the Academic Language domain is distinct in that it is assessed using information and materials from the other four domains. That is, teacher candidates' "skills and abilities" to accommodate their students' varied levels of academic language proficiencies are assessed throughout all stages of the TE.

The TE is a focused learning segment (approximately a week long) that the teacher candidate chooses to teach her or his students in a classroom. A TE typically involves submission of several artifacts, including lesson plans, copies

Task	Domain				
	Planning	Instruction	Assessment	Reflection	Academic Language
1	Items 1, 2, 3				Items 11, 12
2					
3		Items 4, 5			
4			Items 6, 7, 8		
5				Items 9, 10	

Figure 1. Illustration of the structure of the PACT teaching event.
Note. PACT = Performance Assessment for California Teachers.

of instructional and assessment materials, video clips of teaching, a summary of whole class learning, an analysis of student work samples. In addition, the teacher candidate supplies written commentaries in response to item/prompts: these responses describe the teaching context, analyze teaching practices, explain outcomes for students, and provide reflections on lessons learned about one's teaching practice and student learning (PACT, 2009). These submissions result in a full portfolio of the teacher candidate's plans, activities, instructional materials, assessments, and reflections for the learning segment. The portfolios are scored by centrally trained raters generally local to each candidate's institution. Each rater scores all 12 items for a distinct set of teacher candidates on a 4-point scale with Level 1 being the lowest level of performance, indicating a failure to adequately complete the task, and Level 4 being the highest, indicating exceptional performance. A candidate can fail one of two ways (Pecheone & Chung, 2007): If the candidate receives more than one score of "1" for the questions for a particular domain; or if the candidate has more than 3 scores of "1" across all 12 questions. Candidates who fail or who are near failing are double-scored. A random sample of TEs is double-scored to check for rater consistency according to the instrument developers (PACT, n.d.).

Previous Validity and Reliability Studies of the PACT

The most comprehensive validation study prepared by the instrument developers is based on pilot PACT data. The PACT instrument was initially piloted in 2002-2003 and 2003-2004; it was further piloted until 2007. It was approved for operational use in 2008 after the publication of the comprehensive study conducted by a research group led by the PACT consortium (Pecheone & Chung, 2007). This initial technical report was produced from 625 of the 700 completed TEs in the 2003-2004 pilot administration of the PACT. The score data from TEs included subject areas such as Elementary Literacy and Mathematics.

Using the *Testing Standards* (AERA, APA, & NCME, 1985, 1999) for guidance, Pecheone and Chung (2007) investigated a host of issues, including content validity, bias and fairness, construct validity, criterion-related concurrent validity, and score consistency and reliability on pilot data. Generally, their findings were positive. For instance, they provided supporting evidence for content validity by aligning the TE tasks with the California Teaching Performance Expectations. In their bias (i.e., impact) review, they found no significant differences between scores by candidates' race/ethnicity, percentage of English Language Learner students in their classrooms, grade level taught, students' academic achievement level, or months of previous paid teaching experience. They did, however, find that, on average, females scored significantly higher than males and that candidates teaching in high socioeconomic/suburban schools scored higher than candidates teaching in low socioeconomic or urban/inner-city schools. An exploratory factor analysis of the pilot data from the Elementary Literacy Teaching Event found two distinct factors—one for Planning, Instruction, and Academic Language and another for Assessment and Reflection—indicating that the PACT instrument is tapping into distinct constructs of teaching "skills and abilities."[3]

Other studies have focused on aspects of rater reliability evidence. For instance, Porter (2010) examined interrater reliability of scores on PACT TEs completed by teacher candidates in Spring 2009 at a particular California State University. The analysis only included the 23% of the 181 teacher candidates (n = 41) who were double-scored (e.g., scored by two different raters). Porter assessed the interrater reliability with a number of indices and found poor to moderate evidence for consistency in rater scores.

Several quasi-validation studies have used candidate responses to the PACT TE in specific or holistic evaluations of teacher candidates' performance. For instance, Bunch, Aguirre, and Téllez (2009) conducted a qualitative case study using submitted materials for eight teacher candidates' Elementary Mathematics (TE) Teaching Events to further assess their "abilities" in the domain of Academic Language, which the authors defined as the ability to teach and meet the needs of linguistically diverse students in both their academic and English language vocabulary. They found the PACT TE tasks provided useful information in evaluating teacher candidates on this important, but often overlooked, teaching skill.

Table 1. Summary Statistics by Item.

		Items by domain											
		Planning			Instruction		Assessment			Reflection		Academic language	
Time	Statistic	P1	P2	P3	I4	I5	A6	A7	A8	R9	R10	AL11	AL12
2008-2009	*n*	407	407	406	407	407	407	407	311	407	407	407	407
	M	2.74	2.74	2.59	2.54	2.45	2.63	2.44	2.54	2.55	2.54	2.18	2.52
	SD	0.72	0.76	0.71	0.74	0.74	0.78	0.76	0.81	0.71	0.78	0.71	0.67
2009-2010	*n*	1,304	1,303	1,304	1,304	1,304	1,304	1,304	1,297	1,303	1,304	1,304	1,303
	M	2.83	2.82	2.73	2.56	2.43	2.61	2.47	2.56	2.50	2.51	2.27	2.46
	SD	0.65	0.70	0.68	0.71	0.71	0.75	0.75	0.78	0.71	0.74	0.67	0.62
Overall	*n*	1,711	1,710	1,710	1,711	1,711	1,711	1,711	1,608	1,710	1,711	1,711	1,710
	M	2.81	2.80	2.70	2.56	2.43	2.61	2.46	2.56	2.51	2.52	2.25	2.48
	SD	0.66	0.71	0.69	0.71	0.72	0.76	0.75	0.78	0.71	0.75	0.68	0.63

Sandholtz and Shea (2012) explored the relationship between supervisors' predictions and candidates' performance on PACT. Their findings indicate that university supervisors' perspectives did not always correspond with outcomes on the performance assessment, particularly for high and low performers. Though this result might be expected given the greater variability and lower reliability of scores at the ends of the distribution, the study represents an effort to triangulate among different data sources. Along similar lines, Darling-Hammond, Newton, and Wei (2010) have argued that the PACT should be used in concert with several other measures of student teacher learning, as the score data alone are not necessarily sufficient to help guide decisions related to program improvement.

Although these past studies address important aspects of the validity of interpretations from the TE (and the presumed reliability of scores), they do not focus on internal structure validity evidence (i.e., the structure of the performances within and across tasks). The Pecheone and Chung (2007) study is the most comprehensive, but it is based on 2003-2004 pilot data when stakes were low for teachers as their licensure did not depend on their test results. In addition, in subsequent PACT administrations the wording of the instructions, tasks and rubrics have changed along with the number of items for each domain. These qualitative changes potentially threaten the internal structure validity arguments advanced in prior studies.

To warrant ongoing use of the instrument, further standards-based (e.g., AERA, APA, & NCME, 1999) validation studies need to be conducted, for example, by investigating the measurement properties of the PACT instrument under operational testing conditions at scale. The number of institutions participating in the PACT consortium is substantial, and similar alternative performance assessments for teacher candidates are being replicated across the nation (American Association of Colleges of Teacher Education, 2014). Therefore, further study of the meaningfulness and dependability of the scores derived from these types of teacher licensure exams is both warranted and potentially useful to stakeholders—within California and across other states—that intend to adopt instruments that look similar to the PACT.

Data Sample

The data set of teacher candidate PACT scores used for this study was drawn from seven campuses in the University of California and California State University systems that consented to share their data with us. In particular, this data set includes item-level PACT scores for the Elementary Literacy Teaching Event, which along with the Elementary Mathematics Teaching Event, teacher candidates must complete to gain a "Multiple Subjects" (i.e., elementary school teaching) credential. The data set contains first-attempt scores for a total of 1,711 candidates with 407 completing the TE in the 2008-2009 academic year—the first year that California programs were required to implement a teacher performance assessment operationally—and 1,304 in 2009-2010. No identifiers or characteristics were provided at the examinee, rater, or institutional level in accordance with the scope of consent obtained for this study.[4]

Table 1 provides summary statistics by item across all the institutions for each administration year and overall. The mean item scores range from about 2.2 to 2.8 with the Planning items as the easiest and the Academic Language Item 11 (AL11) as the most difficult at both time points.[5] For all of the items, the majority of the scores are 2s or 3s. Approximately 1% to 9% of the item scores are 1s and 5% to 16% are 4s. Generally, there are complete data for all items with the exception of 103 missing scores for the Assessment Item 8 (A8), particularly during the 2008-2009 administration, which mostly occurred for examinees at a single campus. However, no information was available in the data set to explain these missing data.

Statistical Procedures

Item response modeling provides a powerful quantitative method for investigating the structure of latent variables using observed item responses. In technical reports pertaining to Tier I teacher licensure, the psychological construct of "readiness-to-teach" is often described as a composite of an individual's "skills," "abilities," and presumably other "proficiencies." The PACT instrument purports to measure persons with a set of items/tasks on the latent variables mentioned above (i.e., the domains).

Item Response Theory (IRT) models are often used in scoring and scaling educational assessments. However, IRT models are not used in scaling the PACT TEs; rather, observed rater scores based on item rubrics are used. IRT procedures are advantageous in that they can identify empirical patterns of examinee proficiency ("teacher candidate ability") and item ("performance task") difficulty on a single scale or multiple scales. Examining the score results derived from licensure exams such as the PACT, empirical results ("locations") on a continuum ("scale") can then be compared with theoretical expectations of researchers and stakeholders who argue for the meaningfulness of preservice teachers' scores. Specifically, we employ IRT analyses to explore the instrument's internal structure as it pertains to the construct validity evidence more generally (AERA, APA, & NCME, 1999). Through the application of these statistical methods, we investigate the strength of the claims that the PACT scores used to characterize performance are reliable and valid for the intended uses.

The item response model applied in this study is the Partial Credit Model (PCM), which models item responses to polytomous items: PACT tasks are polytomous because they have multiple score outcomes on a rubric. The PCM is the polytomous version of the simple Rasch model for dichotomous items. It models the probability of going from score level j to $j + 1$, such as 2 to 3, in terms of a parameter for the teacher candidate ability (usually represented as a θ), and parameters that describe the difficulty of each relative step in the polytomous score (usually represented as a δ). For the PACT instrument, there are four levels and hence three-step parameters to be estimated for each item. Both the unidimensional form of the PCM, where teacher candidate ability (θ) is a single dimension, and the multidimensional form, where, teacher candidate ability is represented by more than one dimension, are used in this study. Formally, for the unidimensional PCM, the probability that examinee n completes step x (where $x = 1, 2, 3$) for item i is

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} \left(\theta_n - \delta_{ij} \right)}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} \left(\theta_n - \delta_{ij} \right)}, \qquad (1)$$

where θ_n is person n's ability parameter, and δ_{ij} is the step parameter for the jth step for item i (Wright & Masters, 1982). This model expresses the probability of success as a function of the *difference* between the person location (θ_n) and the item-step location (δ_{ij}). For the multidimensional models, a similar formula holds, but it involves a vector of teacher abilities, and also models the correlation between the different dimensions (Adams, Wilson, & Wang, 1997).

All measurement models were run using the psychometric computer program ConQuest (Adams, Wu, & Wilson, 2012). All data handling and subsequent analyses were conducted using R (R Development Core Team, 2012).

A Unidimensional Analysis of the PACT: Fitting a Rasch Measurement Model to the Data

Our first research question addresses the internal structure validity evidence for the use of the summative PACT scores by determining the extent that the item responses to tasks and the teacher candidates' "skills and abilities" can be modeled employing well-established item response measurement models. The advantages of these scaling models for examining the internal structure of a latent variable are numerous, including checking the match between theoretical expectations of the instrument developers and empirical locations of items and persons (Wilson, 2005). We address three issues related to the proposed unidimensional structure of the PACT: model fit, consistency of the results with expectations, and internal consistency reliability. The following subsections describe the procedures we used in the unidimensional scaling analysis of the PACT.

Model fit. We fit the unidimensional PCM to the full data set (n = 1,711) with different difficulty estimates for Items P3 and AL11 in 2008-2009 and 2009-2010 given preliminary differential item functioning (DIF) by time analysis and qualitative investigation of the rubrics. Before interpreting the parameter estimates, it is important to evaluate how well the model fits the full data set.

One way to assess measurement model fit is through an analysis of item fit statistics. Wilson (2005) recommended using weighted mean square fit statistics for the items to detect deviations from model fit. These item fit statistics compare the variability of the residuals, or differences between observed and expected scores for the item of interest and each respondent, over the expected variability assuming the model fits. The recommended tolerance bounds for these fit statistics are between 0.75 and 1.33 (Wilson, 2005). In this study, the item fit statistics for the overall item difficulties are all within these bounds, which means there is evidence for reasonably good measurement model fit to support the unidimensional PCM. The PACT rubrics, in this sense, work reasonably well to form a single scale—from less to more "skills and abilities"—across the four score categories.

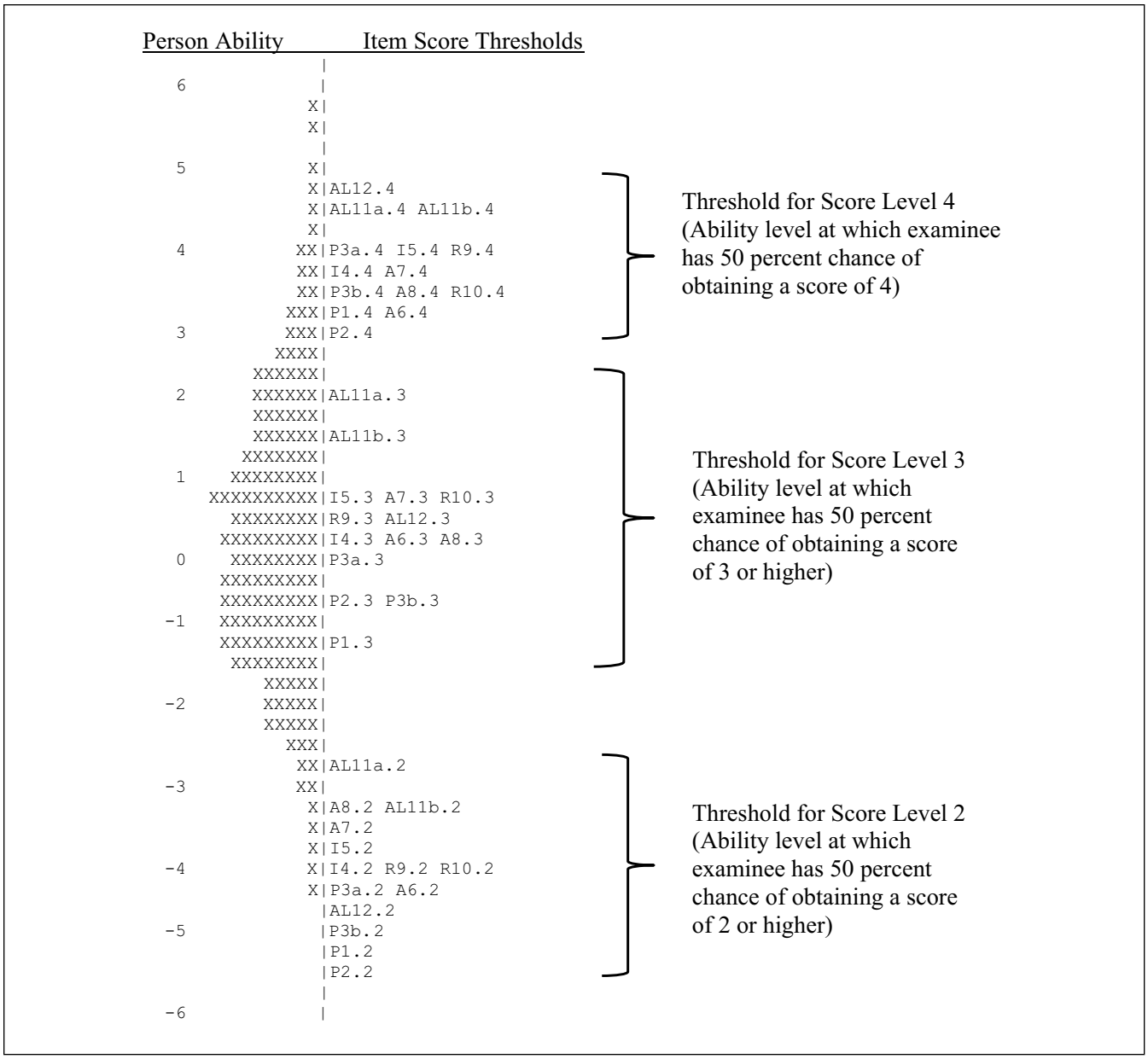

Figure 2. A Wright map for the PACT item score thresholds and person ability estimates using the unidimensional partial credit model. *Note.* Each "X" in the person ability distribution given on the left represents approximately 10 teacher candidates. PACT = Performance Assessment for California Teachers.

Consistency of model results with expectations. An advantage to fitting a PCM to these multiresponse items, as opposed to some of the other polytomous item response models, is that it allows us to construct a Wright map that places the estimated teacher candidate abilities and item-step difficulties on the same scale (Wilson, 2005). This map facilitates inspection of the ordering of the item-step difficulties and facilitates comparison of the standing of the teacher candidates to these item-step difficulties. Accordingly, we can straightforwardly determine whether these distributions are consistent with the instrument developer expectations—expectations, such as, adequate spread of examinee ability and item-step difficulties with consistent ordering of steps. To the extent there is a match between the internal structure posited by the instrument developers and those found by the empirical score analysis, we can add weight to the claim that the PACT measures what it purports to measure.

Figure 2 displays the Wright map with the teacher candidate ability distribution (n = 1,711) shown on the left and the locations of the Thurstonian thresholds for the items on the right. The Thurstonian threshold directly relates the item category difficulties to examinee's ability. Specifically, the

Thurstonian threshold for item score level *k* reflects the location on the latent ability at which the probability of scoring at or above level *k* is .50 (Adams et al., 2012). The Wright map starts with thresholds for scores of 2 as examinees have a 100% chance of earning a score of 1 or higher on any given item. For example, this map shows that item thresholds I4.2, R9.2, and R10.2 are all at −4 on the ability scale, meaning that teacher candidates with PACT ability estimates of −4 have a 50% chance of scoring at Level 2 or higher on the I4, R9, and R10 items. They have less than a 50% chance of scoring at Level 2 for the items with Level 2 thresholds greater than negative 4 (Items I5, A7, A8, and AL11) and of scoring at Level 3 or higher for all items.

The Wright map shows clear "banding" of the 12 items by their score thresholds—All the Level 2 thresholds are clustered together as are the Level 3 and Level 4 thresholds. This banding indicates, for example, that the difficulty across the items of achieving a score of 2 is consistently easier than achieving a score of 3 or 4. In other words, as expected from the PACT rubric descriptions and training protocols, a Level 1 score is similar in difficulty across all items, as are Level 2, 3, and 4 scores. Accordingly, the item-step bandings provide internal structure validity evidence, in part, by demonstrating the expected use of each score category across the operational data set. Put simply, there is evidence to support the claim of meaningfulness and consistent use of the 4 score categories across the structure of the PACT.

Reliability of score results. Reliability, or consistency of scores, is a measure often used in evaluating the utility of a test. High reliability of teacher candidate scores is an expected feature of the PACT instrument. Decision makers need to know that whatever the PACT purports to measure, it does so reliably. While there are several types of reliability indices (e.g., test–retest, alternate forms, rater), we examined the internal consistency indicator generated by the PCM model. The internal consistency coefficient estimates the proportion of variance accounted for by the estimator of a teacher candidates' ability location. This "variance explained" formulation is familiar to many through its use in analysis of variance and regression methods. We use it as a basis for calculating the *separation reliability* (Wright & Masters, 1982, p. 106) which is an IRT equivalent of Cronbach's alpha. The person-separation reliability index of .92 that we obtained for the PACT instrument is high and indicates good internal consistency.[6]

The Wright map indicates that the item score Thurstonian thresholds cover the entire range of teacher candidates' ability estimates, indicating that the PACT instrument can discriminate among teaching "skills and abilities" to the full extent of the range on the scale. This "coverage" is related to the reliability of the PACT, as the better the match between item and teacher location, the better the reliability tends to be (Wilson, 2005). The Wright map also shows that most examinees score at least a 2 on the items as most of the ability estimates fall above the items' Level 2 thresholds. We further note that for teacher candidates with the highest proficiency on the scale, scoring at Level 4 on items remains difficult to achieve, which we also observed directly in the data with few teacher candidates receiving scores of 4 on the items.

A Multidimensional Analysis of the PACT

The content validity argument advanced by the PACT instrument developers identifies five teaching practice domains by groupings of ratings of responses to four tasks. According to the technical report, a total of 12 rubrics are used to evaluate performance in the five domains. Implicitly, these tasks, rubrics, and domains are multidimensional:

> 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. (Pecheone and Chung, 2007, pp. 25-27, emphasis added)

We investigate these claims by determining how well various multidimensional IRT models fit the operational data. In particular, we assess the fit and utility of the unidimensional model, a model based on the tasks, a model based on the domains, and other models driven by empirical findings. The subsections describe the procedures we used in the multidimensional analyses of the PACT.

The unidimensional model. The unidimensional model fit in the previous section provides a single ability estimate for each teacher candidate. However, the TE is scored on five different domains and five tasks as shown in Figure 1 with Planning, Instruction, Assessment, and Reflection items representing their own tasks and Academic Language items spanning across all tasks. Although the unidimensional model fits reasonably well, as noted above in the "Model Fit" section, and provides useful information about some aspects of the internal structure of the PACT TE (such as the uniformity in category difficulties across the items), it gives just one composite estimate for each teacher candidate and hence does not provide explicit measures on the different teacher candidate "skills and abilities" on different aspects of the content (such as the tasks and domains) embodied in the TE.

The unidimensional model is useful in producing ability estimates that can be used to make decisions about overall teacher candidate "readiness-to-teach" at the Tier I licensure level, but if teacher programs or teacher candidates themselves would like more specific information about a candidate's strengths and weaknesses, a multidimensional item

response model would be more informative for teacher educators and program administrators. Multidimensional item response models explicitly model teacher candidate proficiency on multiple latent dimensions. That is, unlike the unidimensional PCM, multidimensional models assign teacher candidates ability estimates on more than one construct.

It may seem contradictory to seek both a unidimensional model and multidimensional models for the PACT, but indeed, that is what is implicit in the provision of both a composite score and subscores by the PACT administrators. In fact, where the multiple dimensions are all moderately to highly correlated, both perspectives can be of practical help in using the results.

The task-based model. We first assessed the multidimensional structure of the Elementary Literacy Teaching Event with a model that parallels the TE's task scoring structure as illustrated in Figure 1. This is a "within-item" multidimensional model as it allows some items to contribute to more than one latent dimension (Adams et al., 1997). As shown in Figure 3a, this task-based model has four dimensions corresponding to the tasks. The Planning, Instruction, Assessment, and Reflection items each map onto different task dimensions, but the Academic Language items map onto all of the task dimensions. Although initially one might think that this model should fit well as it follows the intended structure of the TE, in fact, it resulted in relatively poor model fit.

To assess relative model fit, we use a statistical model selection criterion called the Akaike Information Criterion (AIC). By comparing the AIC from this task-based multidimensional model to the AIC of the unidimensional model described in the previous section, we can determine which model fits better. This task-based multidimensional model resulted in an AIC of 34,531 compared with the unidimensional PCM's AIC of 33,587. Smaller AICs indicate better fit, and in this case, the task-based multidimensional model has a considerably higher AIC (difference = 944), surprisingly implying worse model fit than the simpler, unidimensional model. Moreover, some of the individual item (weighted mean square) fit statistics were outside of the acceptable bounds of .75 to 1.33. In particular, the Academic Language items had high item fit statistics (approximately between 1.4 and 2.0), which are often indicative of model misfit (Adams & Khoo, 1996). This result indicates that although the items were designed according to Figure 1, the real teacher candidate data are not consistent with the test structure in Figure 1.

The misfit of the Academic Language items for the task-based multidimensional model might suggest that when the raters used all of the materials across all the tasks to score these items as instructed, the resulting patterns of data are inconsistent with an item-based dimensional model (Wihardini, Castellano, & Wilson, 2013). A different interpretation might be that, as a relatively new or conceptually difficult construct, the Academic Language items may make additional demands on teacher candidates, raters, and/or preparation programs, and this too has made it difficult to fit the data with a task-based dimensional model.

The domain-based model. The misfit of the Academic Language items for the task-based multidimensional model suggests that the Academic Language items may, in fact, represent their own dimension, as may the items from the other domains. Accordingly, we fit a five-dimensional (5D) between-item model with items for each domain mapping onto their own dimension as illustrated in Figure 3b, and, as such, we refer to it as "the domain-based" model of the PACT TE. Unlike the task-based model, this (between-item) multidimensional model maps each item to only one dimension. This model showed good model fit compared with both the task-based model and the unidimensional model (e.g., AIC_{5D} = 33,059 vs. AIC_{1D} = 33,587). In addition, the individual item fits were all reasonably good.

However, in addition to model fit, an important consideration when selecting a multidimensional model is the correlations among the dimensions. If the correlations are very high, then the dimensions are not distinct. The correlations among the dimensions estimated by the domain-based 5D model are the disattenuated correlations, or correlations corrected for measurement error, among the five domains. These disattenuated correlations are given below the diagonal in Table 2, and the correlations among the mean domain scores—not corrected for measurement error—are above the diagonal. As shown in Table 2, the disattenuated correlations range from .79 to .95, which suggests that some of the dimensions are very similar. The strongest pairwise correlations are among the Assessment, Reflection, and Academic Language (ranging from .90 to .95). Such high correlations provide strong evidence that these three domains are not providing information on three distinct teaching "skills and abilities"; they are tapping into a common construct. Given that candidates are rank ordered similarly on Assessment, Reflection, and Academic Language, we can say that knowing a candidate's score on one of the three domains is sufficient to know her or his performance on the other two. This is not to say that the PACT could not be altered so that it supported interpretations on the intended five domains, but rather that as it is currently and given our data sample, we do not find strong evidence that it does indeed support reporting of five distinct domains.

Domain-based models based on emerging hypotheses. Although the domain-based model fits the data well, based on our program experience with PACT candidates, instructors, and raters, we hypothesized that other domain-based models (with fewer than five dimensions) might do a relatively better job of explaining the internal structure of the PACT operational data as demonstrated by the relationship among the domain rubric scores. To examine our hypotheses based on experience with PACT implementation, we examined the broader

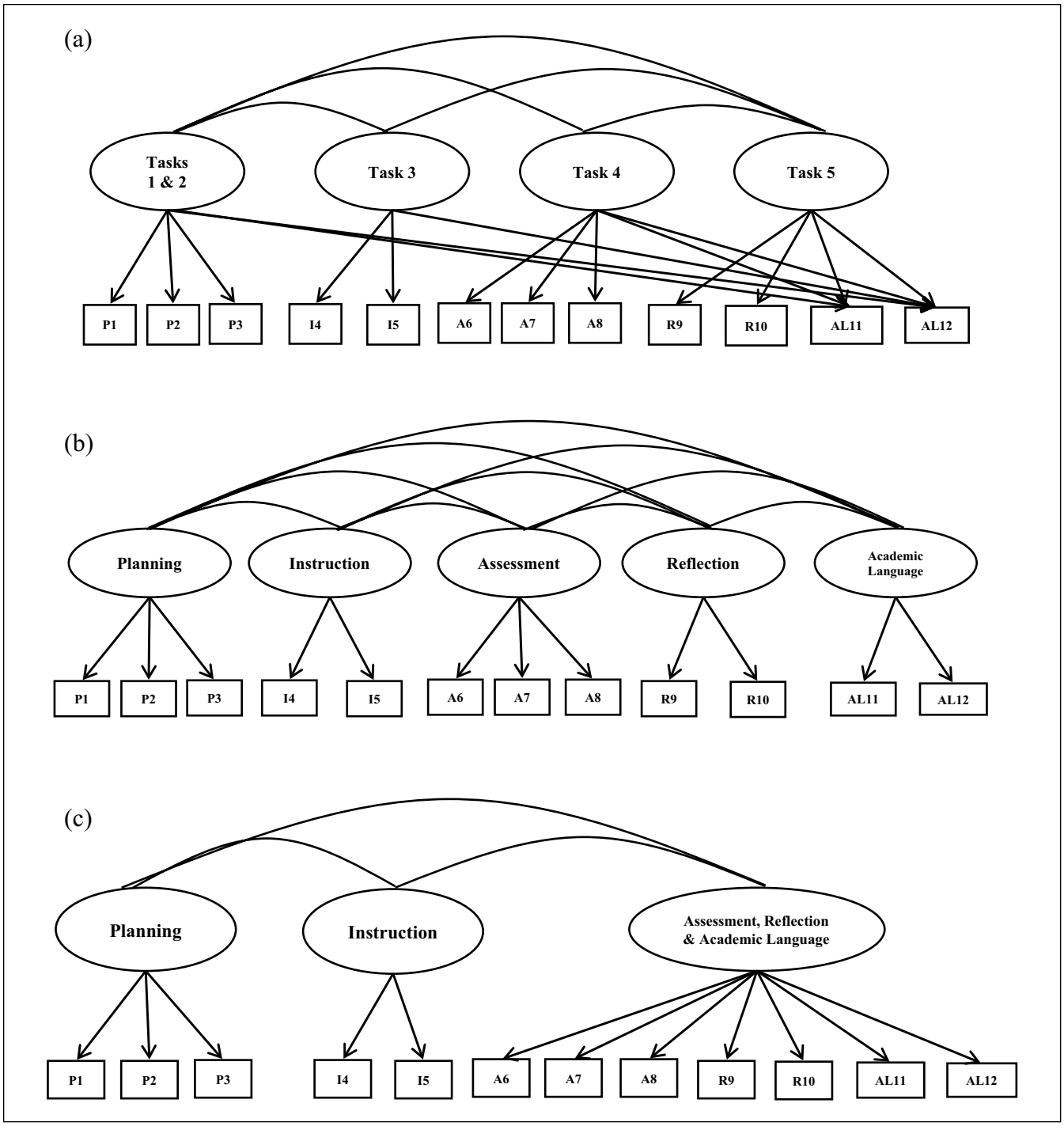

Figure 3. Illustration of multidimensional models defined by structure of the PACT Teaching Event: (a) task-based model, (b) five-dimensional domain-based model, and (c) three-dimensional modified domain-based model with the first dimension defined by Planning items, the second by Instruction items and the third by Assessment, Reflection, and Academic Language (Metacognition) items.
Note. PACT = Performance Assessment for California Teachers.

statistical evidence for an internal structure of PACT scores not fully anticipated by PACT instrument developers. Specifically, we found high correlations among the dimensions in the 5D model, which led us to suspect that fewer constructs may sufficiently describe the various aspects of preservice teacher "ability and skill" purportedly measured by the PACT.

First, as seen in Table 2, the strongest pairwise correlations for the 5D model are among the Assessment, Reflection, and Academic Language (ranging from .90 to .95), suggesting

they are all tapping into a common teacher candidate ability. These three dimensions may be collapsed into one dimension (a latent variable we see as related to metacognition) without large loss of information generated by this particular instrument.

Second, we draw further theoretical support for collapsing the Assessment, Academic Language, and Reflection domains of the PACT into a single dimension of teacher candidate "skills and abilities" from the literature on situated inquiry, modeling in learning theory, and metacognition (White & Frederiksen, 1998).[7] Research on metacognition, in particular, suggests that differences between teacher candidates may be substantially related to how they think about their own thinking, in situations they experience (Dewey, 1910). Thinking about one's teaching practice reflectively is likely to be highly constrained by the learning environment itself; in this case, the cognitive and situational demands of both the teaching placement in which the TE occurs and the teacher preparation environment more generally.

Based on informal exit questioning at our local preparation programs, teacher candidates report that the act of reflection occurs for the most part at the end of the PACT performance cycle (i.e., end of final semester). For many, a "cramming" effect occurs close to the deadline as the teacher candidates write up responses to the multitude of task prompts (J. Lovell, personal communication, April 20, 2013). Compounding the extraneous "noise" (e.g., anxiety over time management, document formatting, use of genre-specific writing skills) associated with performance assessments, the substantive meaning of these events, artifacts, and actions must later be recalled by teacher candidates, which for many feels more like an "open book" written test by the end of the cycle rather than an authentic "live" commentary on their pedagogy.

An important rationale for adopting the three-dimensional model is that it yields a set of score interpretations that account for both quantitative statistical findings and "local" qualitative observations about candidate practices. For many teacher candidates, metacognitive "overload" may in fact constrain their "skills and abilities" to carefully assess student results, to explain students' academic language needs, and to provide genuine research-based reflections about "next steps" in teaching future lessons related to the TE being described. Candidates being assessed by the PACT instrument may literally use the metacognitive dimension as an *afterthought on practice* (long since enacted) particularly as they struggle to pull together written work in a coherent, well-organized, persuasive tone that will likely convince an audience (the rater). In other words, the "skills and abilities" in the three PACT content domains (Assessment, Reflection, and Academic Language) are better represented as a single meta-cognition dimension, one that calls upon teacher candidates' use of general skills (e.g., to recall, reflect, rethink, and revise in writing their observations about practice). We argue that valid score interpretation for these three domains leans toward a more general claim about the metacognitive dimension, that is, the teacher candidates' use of a variety of "skills and abilities" that demonstrate how well s/he can persuasively write toward task prompts to meet a deadline at the end of their PACT TE performance cycle.[8]

There is another plausible explanation which could account for the close association between the subscores on these three domains. We hypothesize that there is a substantial degree of confusion about the Academic Language domain despite efforts of the instrument developers to add scaffolds to the PACT Handbook. Informal self-reports from candidates, instructors, and raters continue to emphasize the difficulty with making sense of prompts, rubrics, and the PACT handbook—particularly in the domain of academic language.[9] Wihardini et al. (2013), in a qualitative study parallel to this one, found that instructors and raters struggle with interpreting task prompts, rubric descriptors, and the footnotes inside the rubrics.

To investigate our hypotheses about the role of metacognition and situational learning constraints, we fit the three-dimensional (3D) modified domain-based model to the data set as shown in Figure 3c. In this approach, three Wright Maps, like the one provided for the unidimensional model, can be used to better understand the structure of the PACT instrument. Based on this model, there are three proficiency scales related to the PACT: Planning, Instruction, and Metacognition which comprises Assessment, Reflection, and Academic Language as a single dimension. Figure 4 shows the three Wright Maps for the three dimensions. This modified domain-based 3D model has good item fit, but does not exhibit as good model fit as the full domain-based 5D model (AIC_{3D} = 33,130 vs. AIC_{5D} = 33,059). However, the three dimensions in the 3D model represent more distinct aspects of teacher candidate readiness and are thus more informative than the 5D model. Moreover, the dimension person-separation reliabilities are slightly higher for the 3D model than the 5D model: For the 5D model, the reliabilities are .86 for Planning, .83 for Instruction, .87 for Assessment, .88 for Reflection, and .88 for Academic Language; for the 3D model the reliabilities are .89 for Planning, .85 for Instruction, and .92 for the combined Assessment, Reflection, and Academic Language dimension.

For the modified domain-based 3D model, the Assessment-Reflection-Academic Language (what we call the "Metacognition") dimension is correlated .849 with the Planning dimension and .852 with the Instruction dimension, and these two dimensions are correlated .835 with each other. These three separate dimensions are tapping into constructs of teaching "skills and abilities" that seem to be more distinct than the five constructs in the 5D domain-based model, which, as seen in Table 2, has three correlations greater than or equal to .90. Thus, this 3D model provides internal structure validity evidence that the Elementary Literacy Teaching Event is assessing different aspects of the candidates' "skills and abilities" (in the dimensions of Planning, Instruction,

Table 2. Observed Correlations Between Mean Domain Scores (Above Diagonal) and Disattenuated Correlations Between Domains/Dimensions (Below Diagonal).

	Planning	Instruction	Assessment	Reflection	Academic language
Planning		0.64	0.64	0.63	0.65
Instruction	0.81		0.61	0.62	0.61
Assessment	0.80	0.79		0.70	0.67
Reflection	0.82	0.84	0.92		0.67
Academic Language	0.84	0.84	0.90	0.95	

and Metacognition), but not necessarily in construct-relevant ways anticipated by the PACT instrument designers.

Discussion

Analyzing score data from a large sample of teacher candidate responses across two public California university systems, we found a sufficient degree of internal structure validity evidence to support the continued use of the PACT instrument as intended to measure California teacher candidates' "skills and abilities" in accordance with the state's professional standards in teaching. Our quantitative study of the Elementary Literacy Teaching Event reveals that item responses and teacher candidate proficiencies can be modeled employing well-established item response measurement models, which yield useful information for more valid score interpretation.

The first research question in this validation study examined internal structure validity evidence by determining the extent that the item responses and teacher candidate proficiencies can be modeled using well-established item response measurement models. In general, we found that the unidimensional PCM fit the data reasonably well, resulted in a high reliability estimate, and offered a useful tool to relate teacher candidate PACT ability estimates with item score thresholds. As seen in the Wright Map in Figure 2, teacher candidate ability estimates spanned the range of the item score thresholds and these thresholds were grouped together by score level. Thus, the scoring rubrics were used consistently across items. In terms of internal consistency, the PACT instrument's reliability is high ($r = .92$).

Nonetheless, absent any absolute standards for reliability in teacher licensure regimes, a better approach to establishing an acceptable r may be to consider each type of application individually and develop specific standards based on the context for use. For example, if a teacher performance assessment instrument such as the PACT is used to make a single division into two groups ("pass/fail"), then a reliability coefficient may be quite misleading, using, as it does, data from the entire spectrum of the respondent locations. In such a situation, it may be better to investigate false positive and false negative rates in a region near the cut location (Wilson, 2005). In its current form, the PACT scores cover the entire range of the 4-point scale and make moderate use of the extremes (1s and 4s).

The second research question in this validation study investigated potential dimensions or constructs embedded in the elementary PACT TE. While the unidimensional model fits the data reasonably well and provides useful information about the internal structure of the PACT TE for licensure, it does not provide any information about teacher candidate "skills," "abilities," and "proficiencies" on different aspects of the underlying constructs.

Beyond reporting a global (pass/fail) score for teacher candidates, many stakeholders are interested in dimensions of performance captured by the PACT instrument. Some hope the data can be used formatively, while others seem more interested in summative uses for reporting. Accordingly, we examined which multidimensional models fit the operational data sample better. First, we found that the task-based multidimensional model resulted in relatively poor model fit, particularly with respect to Academic Language. The misfit of the Academic Language items may mean that these items tap into their own unique dimension in the Elementary Teaching Event; or alternatively, the item prompts, task descriptions, rubrics, and so forth may be too unstable at this point in the implementation process to warrant meaningful interpretation.

Given the misfit of the task-based model and theoretical considerations, we fit a full domain-based 5D model and then a modified domain-based 3D model to the operational data set. We found that, although the 5D model fit relatively better than the 3D model, it yielded dimensions with very high correlations, which makes interpretation difficult. Hence, we concluded that the modified domain-based 3D model was the most informative with distinct dimensions for Planning, Instruction, and Metacognition (which included Assessment, Reflection, and Academic Language based on correlational findings). This finding differs from Pecheone and Chung's (2007) earlier validation study on 2003-2004 pilot data. They found that two factors best explained the internal structure of the PACT pilot scores: one for Planning, Instruction, and Academic Language and another for Assessment and Reflection.

In both cases, from the data available, it was found that the PACT internal structure validity evidence did not support the content validity argument proposed by the instrument designers for a five domain-based factor structure. We acknowledge that any similarities among particular domain subscores may be due to rater scoring behavior or teacher candidate performance on such domains, or due to the

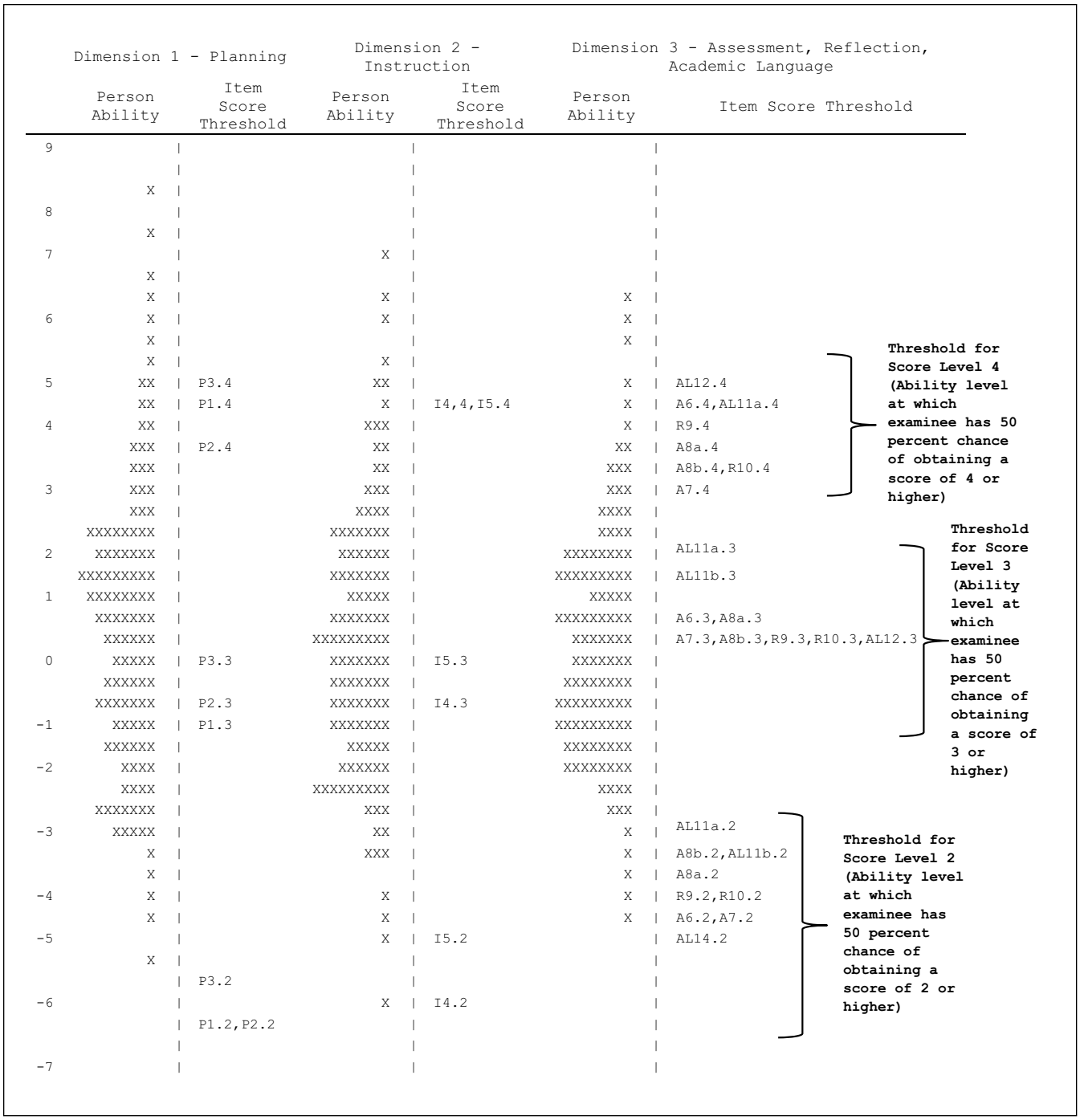

Figure 4. A Wright map for the PACT item score thresholds and person ability estimates using the multidimensional partial credit model.

Note. Each "X" in the person ability distribution given on the left represents approximately 13.9 teacher candidates. PACT = Performance Assessment for California Teachers.

theoretical similarity of the latent constructs, such as Assessment, Reflection, and Academic Language all representing aspects of metacognition. A host of issues could explain the conflicting score information.

Our findings do not deny the possibility of other plausible, evidence-based interpretations of the internal structure of the PACT data. Despite the fact that candidates have similar levels of performance or at least rank ordering on each of these (Assessment, Reflection, and Academic Language) domains, when talking about individual candidate's "skills and abilities" we can be tempted to overinterpret, or worse, misuse these subscores. It seems clear that the raw scores do not tell the whole story. Our point is that if stakeholders in teacher education desire instruments that are

instructionally sensitive (Popham, 2006) and which are assumed to detect distinct, stable measurable skills in the area of academic language or assessment or reflection, more work is to be done. Our study provides evidence that more articulation of the construct definitions, coupled with more nuanced instrument development work is needed to better measure teachers' practices in and for our field.[10]

Future Directions

We began this article by accepting the proposition that, in principle, large-scale performance assessments such as the PACT can help guide decision makers in evaluating who is "in" and who is "out" of the teaching profession at the Tier I licensure level. Our research on the internal structure evidence for (and against) meaningful interpretations of PACT scores is motivated by the fact that the instrument has significant consequences for both teachers and for those who prepare them. Teacher candidates are compelled to take this performance-based instrument (or similar constructed response "open book" tests such as the edTPA) if they hope to obtain a Tier I license in California. Moreover, state policy makers and teacher educators who are "data driven" are compelled to take these results seriously to make better decisions regarding the allocation of resources. Some may be tempted to compare programs and institutions to determine the "value added" of individuals (e.g., faculty and cooperating teachers) with respect to the global and subscore data.[11]

Our goal has been to investigate the validity and reliability evidence for the Elementary Literacy PACT instrument's intended uses based on production data obtained in this study. Along with others, we are acutely aware of the need for multiple measures of preservice teacher "skills," "practices," and "proficiencies" (Darling-Hammond et al., 2010). Evidence from teacher candidates' academic coursework, field placements, supervisor ratings, and so forth are essential to making a judgment about readiness to teach in the California classroom. While these other sources of evidence are necessary, we note that none of these types of instruments used to gather data on teacher candidates are required by law to meet technical criteria for validity and reliability.

The PACT, however, must meet these criteria under operational, not just pilot, conditions to warrant large-scale use and replication. Toward this end, we maintain that further validation studies guided by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) are needed to warrant ongoing uses of the score data within and across different states. While a few studies (e.g., Newton, 2010) have been conducted in Connecticut and California to assess the predictive validity of performance-based assessments for beginning teachers (i.e., whether the assessments predict teacher effectiveness as measured by student achievement), future validation studies might bring about a broader and more compelling evidence base of qualitative and quantitative findings to support, modify, or limit particular aspects of global score and subscore interpretations currently attached to this licensing exam.

This study is relevant to those psychometric researchers and teacher educators who are working together to establish a robust body of validity evidence envisioned by the original PACT consortium, in particular those committed to the stated mission of developing an alternative curriculum-embedded assessment using an "evidence-based system" (Pecheone & Chung, 2007). Testing companies and other third parties should not be solely relied upon to conduct future validation studies to explore fundamental validity and reliability issues for teacher performance licensure and certification "tests" at scale. Rather, the end users (teacher candidates, university instructors, cooperating teachers) also have an interest and responsibility to explore what is and is not working with these licensure instruments.

Toward this end, we recommend a program of future validation studies of teacher performance assessments used for licensure (not only the PACT) that directly address the following strands of *Standards for Educational and Psychological Testing* (1999). The *Testing Standards* can serve as a shared framework for developing and vetting claims, evidence and warrants. We offer a few examples of the research questions that follow from a commitment to the most recent *Testing Standards*:

- Content: What do national content experts say about the "skills," "abilities," and "proficiencies" measured by teacher licensure and certification instruments such as PACT? Is there new research on topics such as academic language or formative assessment that could inform the meaning/use of the current domains? Can new technologies capture more complex performances and hence the richness of multiple constructs than current item designs?
- Response processes: What do cognitive labs/exit surveys with teacher candidates tell us about the language, task, technology, and reading demands of the PACT? Are examinees engaging in construct (ir)relevant behaviors? What are the systematic effects, if any, of "noise" in obtaining the target of inference (i.e., readiness to teach in the California classroom)? Can these qualitative findings be triangulated with quantitative IRT studies of person and item misfit test characteristics?
- Internal structure: Beyond the study of the multiple subject PACT TEs, what is the dimensional nature of the single subject PACT TEs (i.e., math, science, language arts, social science, music, art, foreign language, physical education, and so forth)? Which kind of DIF studies are required? Are there characteristics of teacher candidates, raters, or teacher education programs that influence the performance, scoring, and interpretation of PACT results?
- Relations to external variables: Are there systematic convergent/divergent relationships between the PACT

instrument scores and data from student teacher ratings, principal evaluations, Beginning Teacher Support and Assessment (BTSA) observations, and so forth? Which of these, if any, are predictive of candidates' future performance?

- Consequences: In which ways, if any, is the PACT instrument having systematic, positive, or negative consequences on teaching candidates? Does the instrument affect curriculum, program, and/or field placement decisions for candidates adversely?

Each of these questions is rooted in the search for lines of validity evidence that can support intended, ongoing uses of PACT data to make decisions for teacher licensure while simultaneously addressing the potential unintended, inappropriate uses of score data. SB 2042 in California allows for alternative assessments to the state assessment in as much as they are aligned with California's professional teaching Standards. The PACT, as a summative instrument that yields a pass/fail score for preliminary licensure, has met this requirement. Yet a more productive approach to engaging stakeholders in unpacking the meaning and use of PACT scores in preservice teacher education is to improve our understanding of the limitations of those scores themselves. We need better input and output measures of teacher candidates' "skills," "abilities," and "high-leverage practices" in their preservice programs. The PACT is necessary but not sufficient for the larger research task at hand.

If large-scale performance assessments such as the PACT or edTPA are expected to responsibly guide decision makers in evaluating who is "in" and who is "out" of the profession, then the evidence to warrant those judgments must also be evaluated. Shepard (1997) has warned against the misuse of standardized test data in K-12 settings—to shape and limit students' opportunities for growth—not to mention the denigration of the role of assessment in a learning culture. Berlak (2011) and others who work in postbaccalaureate teacher education settings have posed challenging questions to the PACT consortium: Is it fair for student teachers in California to face a "high stakes test" to graduate given the variation in circumstances, resources, and background in preparation itself? Do preservice preparation programs have a responsibility to "teach to the test" which inevitably crowds out part of the curriculum in order to guarantee results? What exactly is being tested in the PACT—beyond persuasive writing, videography, and document management skills? And, how do those tested PACT outcomes correlate with actual teacher performance on similar dimensions in the future classroom where these student teachers work?

Collaboration between teacher educators and psychometricians from California higher education institutions is a necessary precondition for the ongoing, legitimate, and scientific validation of PACT scores based on first principles. In the opening validity chapter of the *Testing Standards* (AERA, APA, & NCME, 1999), we are reminded that

> validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. *When test scores are used or interpreted in more than one way, each intended interpretation must be validated* . . . As validation proceeds, and new evidence about the meaning of a test's scores becomes available, *revisions may be needed in the test, in the conceptual framework that shapes it, and even in the construct underlying the test*. (p. 9, emphasis added)

Any "evidence-based system" of teacher evaluation that promises both formative and summative uses of data demands attention to the concept of validity. As Kane (1994) and other validity experts (Kane, Crooks, & Cohen, 1999) focused on performance assessment remind us, the plausibility of an interpretation depends on evidence supporting the proposed interpretation and on the evidence for refuting competing interpretations. Moreover, we should expect that different types of validity evidence will be relevant to different parts of the argument for use. Kane adds another element of complexity in the validation process with the focus on the kinds of claims instrument developers often make:

> *Claims* that the situations included in licensure examinations are representative of the situations encountered in some area of practice could be supported by expert judgment or by empirical data. *Claims* that the test score is related to some other variable (e.g., current or future performance on certain tasks) can be checked against empirical data on the relationship between the two variables. *Claims* about the internal decision-making processes used by examinees in performing certain tasks may demand experimental studies of task performance for their verification. (p. 136, emphasis added)

To guide the choice and quality of validity evidence gathered for examining particular claims about a score result, we must insist on the use of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999). The current *Testing Standards* call for lines of evidence regarding content coverage, response processes, internal structure, relations to other variables, and consequences. While a validity researcher may decide to focus on claims made at the item or total test score level, she utilizes the *Testing Standards* to evaluate the "network of inferences leading from the [task] score to the statements and decisions" included in the interpretative argument (Kane, 1992). For large-scale assessment programs in teacher licensure and certification such as the PACT, robust technical documentation coupled with a spirit of ongoing, shared public research is needed more than ever to examine claims and decisions that directly impact the quality of the teaching profession.

Appendix

PACT Task Prompts

Table A1. Questions for the PACT Teaching Event.

Domain	Question code	Question text
Planning	P1	How do the plans support student learning of skills and strategies to comprehend and/or compose text?
	P2	How do the plans make the curriculum accessible to the students in the class?
	P3	What opportunities do students have to demonstrate their understanding of the standards/objectives?
Instruction	I4	How does the candidate actively engage students in their own understanding of skills and strategies to comprehend and/or compose text?
	I5	How does the candidate monitor student learning during instruction and respond to student questions, comments, and needs?
Assessment	A6	How does the candidate demonstrate an understanding of student performance with respect to standards/objectives?
	A7	How does the candidate use the analysis of student learning to propose next steps in instruction?
	A8	What is the quality of feedback to students?
Reflection	R9	How does the candidate monitor student learning and make appropriate adjustments in instruction during the learning segment?
	R10	How does the candidate use research, theory, and reflections on teaching and learning to guide practice?
Academic Language	AL11(a)[a]	How does the candidate describe the language demands of the learning tasks and assessments in relation to student language development?
	AL11(b)[a]	How does the candidate describe the language demands of the learning tasks and assessments in relation to students at different levels of English language proficiency?
	AL12	How do the candidate's planning, instruction, and assessment support academic language development?

Note. AL11(a) represents the item in 2008-2009 and AL11(b) represents the item in 2009-2010.
[a]The AL11 item changed from the 2008-2009 academic year to the 2009-2010 academic year.

Authors' Note

The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B110017 to University of California, Berkeley.

Notes

1. Darling-Hammond (2001, 2010) has argued for standardized performance assessment as a tool to improve teacher education practices, provide evidence of quality teaching to schools and school systems, and create a national licensing system based on professional standards. However, a recent policy analysis by Cochran-Smith, Piazza, and Power (2013) questions whether performance assessments will deliver on their promise. See, for example, Linn and Baker (1996), Baker (1997), Haertel (1999), and Mansvelder-Longayroux, Beijaard, and Verloop (2007) on portfolios, performance assessment and their role in accountability reform more generally. Despite reservations about performance-based reform (Rennert-Ariev, 2008), assessments that make a point of acknowledging the teaching context, such as teaching portfolios (Wolf & Dietz, 1998), have been promoted as more "authentic" assessments than alternatives that rely on traditional testing formats and batteries for nearly 20 years.
2. We do not generalize from this study. Rather, we report on its findings with respect to pilot data examined by the authors. Our study uses production data results from the initial attempt of teacher candidates who completed the elementary language arts Performance Assessment for California Teachers (PACT) instrument from 2008-2010. We review the work of Pecheone and Chung (2007) as it represents one of the most comprehensive validation studies of the PACT in its pilot stages.
3. It is worth noting that these findings about the internal factor structure of the pilot data were not presented in contrast with the initial domain structure presented by the content validity review. As Kane and the authors of the *Testing Standards* might note, different lines of validity evidence were not weighed or sorted into an overall argument in these initial studies.

4. The data used in this study comprise a census of candidates who attempted to complete the PACT at the seven institutions from 2008 to 2010. All of the teacher candidates were enrolled in a postbaccalaureate licensure program or a master's degree program combined with the teaching license. All programs in California are bound to the Teaching Performance Expectations and Standards and thus share the same outcome goals. While programs vary in size and geographical location, the data sample is consistent with the population of public programs across the State.
5. In previous studies (e.g., Duckor, Castellano, Téllez, & Wilson, 2013), we considered the stability of the instrument for the time periods we collected data. In other words, we want to ensure that the wording and structure of the instrument itself is constant over the 2008-2009 and 2009-2010 test administrations. We obtained the relevant archival item/task prompts and rubrics (N. Merino, personal communication, November 21, 2011). Qualitative investigations of the item prompts and rubrics in 2008-2009 and 2009-2010 revealed one significant change in wording for an item prompt and rubric. This change was for the Academic Language domain, specifically Item 11 (AL11). In 2008-2009, question AL11 asked "How does the candidate describe the language demands of the learning tasks and assessments in relation to student language development?", whereas in 2009-2010 it asked "How does the candidate describe the language demands of the learning tasks and assessments in relation to students at different levels of English language proficiency?". Accordingly, the focus of item AL11 shifted from candidates describing language demands for students with any student language development impediment to only English Language Learner (ELL) students. Similarly, the corresponding score level descriptors for the 2009-2010 rubric focus on English Language proficiency, while the 2008-2009 rubric does not mention this specific type of language development. For instance, the Level 2 descriptor in the 2008-2009 rubric expects candidates to identify their students' language demand in different modalities (speaking, listening, reading, and writing), while the respective level in the 2009-2010 rubric refers only to the language demand of ELL students in oral and written tasks. The shifting focus onto ELLs and their learning in the two language modalities are also found in the other levels' descriptor of the latter rubric. The reference solely to ELLs and their speaking and writing capacities in the 2009-2010 rubric changes the focus of the teacher candidates' actions for this question from the previous year's administration, making this item substantively different across time. Consequently, we cannot consider item AL11 as invariant over time and thus we treat item AL11 scored using the 2008-2009 and 2009-2010 rubrics as two separate items in our subsequent analyses.
6. Our study does not report on rater reliability as information on the raters was not available for the scope of consent. Pecheone and Chung (2007) in their analyses of pilot data examined "assessor reliability" on the five domains based on the double-scored candidates' scores; as might be expected, they found an overall interrater reliability of 0.88 for the pilot study data set which is slightly lower than the traditional internal consistency indicators we report here. Interestingly, Standard 3.23 (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) notes that scorer reliability and potential rater drift, particularly as assessment systems move from pilot to implementation phases, warrants continuous evaluation. There have been no follow-up peer-reviewed rater reliability studies on the PACT since this initial pilot study.
7. White and Frederiksen (1998) noted in the context of K-12 education: "The paradox is that, to understand [any] complex activity, one needs to do it, but to do it, one needs to understand it. The instructional solution we developed combines aspects of prior work on metacognition (A. Brown et al., 1983) and situated cognition (J. Brown, Collins, & Duguid, 1989). It scaffolds carefully the inquiry process for students so that they can begin practicing it. This is analogous to the first stage of an apprenticeship in which novice participants enter a community of practice (Lave, 1988)" (p.9). With this "situated" teaching and learning approach in teacher education often comes the problem of how to measure actual teacher candidates whose practices are mediated by their own writing reflections about (and memory of) those practices. Inferences drawn from instruments attempting to capture "authentic" teacher performance (i.e., observed practices from written and video artifacts) which are situated in multiweek, multifaceted contexts can be further confounded by extraneous construct irrelevant "noise" such as latent writing ability, first language facility, computer proficiency, and other *noncognitive* skills related to academic mindset, perserverance and so forth. The question remains whether metacognition and non-cognitive skills are the intended targets of inference for the PACT.
8. As Shulman (1987) wrote, "As we have come to view teaching, it begins with an act of reason, continues with a process of reasoning, culminates in performances of imparting, eliciting, involving, or enticing, and is then thought about some more until the process can begin again. In the discussion of teaching that follows, *we will emphasize teaching as comprehension and reasoning, as transformation and reflection*. This emphasis is justified by the resoluteness with which research and policy have so blatantly ignored those aspects of teaching in the past" (p. 13, emphasis added).
9. The *Testing Standards* (AERA, APA, & NCME, 1999) are clear: "Theoretical and empirical analysis of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of the performance (p. 12) . . . Studies of response processes are not limited to the examinee . . . relevant validity evidence includes the extent to which the processes of observers or judges are consistent with the intended interpretation of scores" (p. 13). Response process validity studies, which include systematic formal exit surveys and/or cognitive labs have *not* been conducted on the PACT.
10. We gratefully acknowledge the kind of reflective approaches to rethinking the PACT instrument design provided by one of our anonymous reviewers: "These rubrics, too, tend to mix domains—intentionally for academic language—but potentially less intentionally for planning and assessment. If I were writing this article, I would recommend that the rubrics be restructured to separate criteria and domains. This would provide for more test items and likely increase reliability based on simpler (and more consistent) decisions and more items. I'll offer an example to show what I mean. EL1 could be subdivided into three criteria:

1. Learning tasks focus on *multiple dimensions* of literacy learning. (I would leave out reference to assessment tasks under instruction, as that confounds the two domains.)
2. There are *clear connections* among facts/conventions/skills, and strategies for comprehending and/or composing text.
3. There is a *progression* of learning tasks that guides students to build understandings of the central literacy focus of the learning segment.

This approach could lead to a test of about 36 items. Three proficiency levels (e.g., demonstrated, partially demonstrated, not demonstrated) would easily work and likely yield a more consistent set of ratings, too. These three categories are more easily differentiated. New cut-score setting would be needed, but it would be worth the effort to get subscales that work."

11. Any warrants for efficacy claims would, at a minimum, have to derive from complex multilevel, multidimensional latent regression analyses that carefully investigate the effect(s) of campus, program, instructor, placement, and rater on teacher candidate ability estimates. The problems of interpreting the PACT or other TPA-like score results in so-called value added analyses are too numerous to address here. From a psychometric perspective, the current construct definitions, instrumentation design, and scaling properties of the PACT cannot plausibly detect individual differences in consistent or meaningful ways at "grain size" to offer stable or appropriate formative feedback to teacher candidates. The current configuration of the Elementary Literacy PACT instrument supports *limited summative*—not formative—uses and interpretations of score data.

References

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Melbourne: Australian Council for Educational Research.

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23.

Adams, R. J., Wu, M., & Wilson, M. (2012). ConQuest 3.0 [Computer program]. Hawthorn: Australian Council for Educational Research.

American Association of Colleges of Teacher Education. (2014). *Teacher performance assessment consortium*. Retrieved from http://edtpa.aacte.org/

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: Author.

Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, *36*(4), 247-254.

Berlak, A. (2011). Can standardized teacher performance assessment identify highly qualified teachers? In R. Ahlquist, P. Gorski, & T. Montano (Eds.), *Assault on kids: How hyper-accountability, corporatization, deficit ideologies, and Ruby Payne are destroying our schools* (pp. 51-62). New York, NY: Peter Lang.

Brown, A., Bransford, J., Ferrara, R., & Campione, J. (1983). Learning, remembering, and understanding. In J. H. Flavell & E. M. Markman (Eds.), *Handbook of child psychology, Vol. 3: Cognitive development* (4th ed., pp. 77-166). New York, NY: Wiley.

Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32-42.

Bunch, G. C., Aguirre, J. M., & Téllez, K. (2009). Beyond the scores: Using candidate responses on high stakes performance assessment to inform teacher preparation for English learners. *Issues in Teacher Education*, *18*(1), 103-128.

California Commission on Teacher Credentialing. (2012). *CalTPA California Teacher Performance Assessment: Implementation manual*. Retrieved from http://www.ctc.ca.gov/educator-prep/TPA-files/CalTPA-implementation-manual.pdf

Cochran-Smith, M. (2005). The new teacher education: For better or worse? *Educational Researcher*, *34*(7), 3-17.

Cochran-Smith, M., Piazza, P., & Power, C. (2013). The politics of accountability: Assessing teacher education in the U.S. *The Education Forum*, *77*(1), 6-27.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Darling-Hammond, L. F. (2001). Portfolio as practice: The narratives of emerging teachers. *Teaching and Teacher Education*, *17*(1), 107-121.

Darling-Hammond, L., Newton, X., & Wei, R. C. (2010). Evaluating teacher education outcomes: A study of the Stanford Teacher Education Programme. *Journal of Education for Teaching*, *36*(4), 369-388.

Dewey, J. (1910). *How we think*. Boston, MA: Heath.

Duckor, B., Castellano, K., Téllez, K., & Wilson, M. (2013, April). *Validating the internal structure of the Performance Assessment for California Teachers (PACT): A multi-dimensional item response model study*. Paper presented at the annual meeting of the American Educational Research Association Conference, San Francisco, CA.

Farkas, S., & Johnson, J. (1997). *Different drummers: How teachers of teachers view public education*. New York, NY: Public Agenda.

Haertel, E. H. (1999). Performance assessment and educational reform. *Phi Delta Kappan*, *80*(9), 662-666.

Haertel, E. H. (2013). Getting the help we need. *Journal of Educational Measurement*, *50*(1), 84-90.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity argument for accountability testing. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (The 104th Yearbook of the National Society for the Study of Education, Part 2) (pp. 1-34). Malden, MA: Blackwell Synergy.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, *112*, 527-535.

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluations & the Health Professions*, *17*, 133-159.

Kane, M. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, *50*(1), 115-122.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17.

Lave, J. (1988). *Through the supermarket: Cognition in practice*. Cambridge, UK: Cambridge University Press.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessments be psychometrically sound? In J. N. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-Fifth Yearbook of the National Society for the Study of Education, Part 1) (pp. 84-103). Chicago, IL: National Society for the Study of Education (distributed by the University of Chicago Press).

Mansvelder-Longayroux, D. D., Beijaard, D., & Verloop, N. (2007). The portfolio as a tool for stimulating reflection by student teachers. *Teaching and Teacher Education*, *23*(1), 47-62.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York, NY: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement*, *50*(1), 91-98.

Murray, F. B. (2001). From consensus standards to evidence of claims: Assessment and accreditation in the case of teacher education. In J. Ratcliff, E., Lubinescu, & M. Gaffney (Eds.), *New directions for higher education: How accreditation influences assessment* (pp. 49-65). San Francisco, CA: Jossey-Bass.

Newton, S. P. (2010). *Pre-service performance assessment and teacher early career effectiveness: Preliminary findings on the Performance Assessment for California Teachers*. Stanford, CA: Stanford Center for Assessment, Learning, and Equity, Stanford University. Retrieved from http://scale.stanford.edu/index.php?option=com_docman&task=cat_view&gid=71&Itemid

Pecheone, R. L., & Chung, R. R. (2007). *Technical report of the Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003-04 pilot year*. PACT Consortium. Retrieved from http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf

Performance Assessment for California Teachers. (n.d.). *Frequently asked questions*. Retrieved from http://www.pacttpa.org/_main/hub.php?pageName=FAQ

Popham, J. (2006, March). A test is a test is a test—Not! *Educational Leadership*, *64*(6), 88-89.

Porter, J. M. (2010). *Performance Assessment for California Teachers (PACT): An evaluation of inter-rater reliability*. University of California, Davis. ProQuest Dissertations and Theses. Retrieved from http://search.proquest.com/docview/757340264?accountid=14496(757340264)

Raths, J., & Lyman, F. (2003). Summative evaluation of student teachers: An enduring problem. *Journal of Teacher Education*, *54*(3), 206-216.

R Development Core Team. (2012). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rennert-Ariev, P. (2008). The Hidden Curriculum of Performance-Based Teacher Education. *Teachers College Record*, *110*(1), 105-138.

Sandholtz, J. H., & Shea, L. M. (2012). Predicting performance: A comparison of university supervisors. *Journal of Teacher Education*, *63*(1), 39-50.

Senate Bill 2042. (1998). Retrieved from http://www.leginfo.ca.gov/pub/97-98/bill/sen/sb_2001-2050/sb_2042_bill_19980918_chaptered.html

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education* (pp. 405-450). Washington, DC: American Educational Research Association.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*, 5-8.

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, *57*(1), 1-22.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, *16*, 3-118.

Wihardini, D., Castellano, K., & Wilson, M. (December, 2013). *Unpacking the difficulty in assessing and responding to Academic Language aspects of the Performance Assessment for California Teachers*. Poster presented at the 92nd California Educational Research Association, Anaheim, CA.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York, NY: Psychology Press.

Wise, A. E., & Leibbrand, J. A. (2001). Standards in the new millennium: Where we are, where we're headed. *Journal of Teacher Education*, *52*, 244-254.

Wolf, K., & Dietz, M. (1998). Teaching portfolios: Purposes and possibilities. *Teacher Education Quarterly*, *25*(1), 9-22.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Author Biographies

Brent Duckor, PhD, is a teacher educator in the Single Subject Credential Program at SJSU's Lurie College of Education. He is a founding faculty member of the EdD Leadership Program. His interest in performance assessments, portfolios, and school-based accountability began in 1996 when he taught at Central Park East Secondary School in New York City.

Katherine E. Castellano, PhD, is an Institute of Education Sciences postdoctoral fellow at the University of California, Berkeley in the Quantitative Methods and Evaluation program. Her research focuses on addressing key educational policy issues with statistical applications.

Kip Téllez, PhD, is a professor and former Chair in the Education Department at the University of California at Santa Cruz. His research interests include teacher education, second language teaching, and the intersection of the two.

Diah Wihardini is a foreign Fulbright scholar from Indonesia. She is currently enrolled in the Quantitative Methods and Evaluation PhD program at UC Berkeley's Graduate School of Education. Her research interests include measurement, policy implications of international assessments, and evaluation of teacher performance.

Mark Wilson, PhD, is a professor of education at the University of California, Berkeley, where he teaches courses on measurement in the social sciences, multidimensional measurement and applied statistics. He was the president of the Psychometric Society for 2011-12, and also became a member of the U.S. National Academy of Education in the same year. He has chaired two US National Research Council committees.