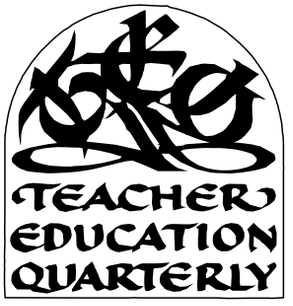


Assessing Academic Language in an Elementary Mathematics Teacher Licensure Exam

By Katherine E. Castellano, Brent Duckor, Diah Wihardini, Kip Telléz, & Mark Wilson

With the adoption by most states of the Common Core State Standards (CCSS) for English language arts and literacy and for mathematics (CCSS Initiative, 2010a, 2010b) comes major changes in public education that will affect instructional practice, curriculum, and assessment across the nation. Heritage, Walqui, and Linquanti (2015) argued that the success of these policy changes will depend, in part, on several important shifts in educators' perspective on language use and language learning, such as from an individual to a socially engaged activity, from a linear process aimed at correctness and fluency to a developmental process on comprehension and communication, and from a separate area of instruction to an embedded component of subject-area activities.

Lee, Quinn, and Valdés (2013) discussed the language learning challenges and opportunities in the new science, math, and language arts standards. They noted that teachers will have to adopt new ways of thinking about teaching and

Katherine E. Castellano is a psychometrician with Educational Testing Service, San Francisco, California. Brent Duckor is an associate professor in the Department of Secondary Education of the Lurie College of Education at San Jose State University, San Jose, California. Diah Wihardini is a graduate student in the Graduate School of Education at the University of California, Berkeley. Kip Téllez is a professor in the Education Department at the University of California, Santa Cruz. Mark Wilson is a professor in the Graduate School of Education at the University of California, Berkeley. KEcastellano@ets.org, Brent.Duckor@sjsu.edu, diah.wihardini@berkeley.edu, ktellez@ucsc.edu, & MarkW@berkeley.edu

learning for all students, particularly English language learners (ELLs), arguing for

> a parallel redefinition of what it means to support learning language in the science classroom by moving away from the traditional emphasis on language structure (phonology, morphology, vocabulary, and syntax) to an emphasis on language use for communication and learning. . . . We propose that when students, especially ELLs, are adequately supported to "do" specific things with language, both science learning and language learning are promoted. . . . Furthermore, [our] conceptualization could be applicable to other subjects, especially CCSS for English language arts and literacy and for mathematics. (pp. 1–2)

Teacher preparation programs play a critical role in the adoption and sustainability of CCSS reforms. In many instances, such programs have anticipated these calls for change by developing the mathematical knowledge base and pedagogical skill set of new elementary school mathematics teachers in their courses and curricula. Building on a firm knowledge base, teacher educators have drawn from key writings by Pimm (1987) and others (e.g., Morgan, 1998; Spanos, Rhodes, Dale, & Crandall, 1988) who have dispelled the view that mathematics is a language-free discipline. Research by MacGregor and Price (1999) found that a general knowledge of syntax in language is associated with mastering the syntax of algebra. Furthermore, Danesi (2003) has demonstrated that knowledge of metaphor is key to understanding and solving "story problems." Yet many elementary school teachers, especially credential candidates themselves, may lack an understanding of the complex relationship between language and mathematics learning. Moreover, programmatic changes are needed in collaborative relationships between English as a second language (ESL) and content teachers regarding disciplinary language use and academic language (Valdés, Kibler, & Walqui, 2014).

This relationship between language and any discipline is generally referred to as *academic language* (AL). Definitions of AL are varied, but a general consensus has emerged (Snow, 2010; Snow & Uccelli, 2009). In Snow's view, AL refers "to the form of language expected in contexts such as the exposition of topics in the school curriculum, making arguments, defending propositions, and synthesizing information" (p. 450), but she has admitted that the boundaries of this definition remain fuzzy. Others have defined AL by pointing out what it is not: AL is "language that stands in contrast to the everyday informal speech that students use outside the classroom environment" (Bailey & Butler, 2003, p. 9). Still others have suggested that it is defined by its use: AL is needed for "tasks that language users must be able to perform in the content areas" (Chamot & O'Malley, 1994, p. 40). Bunch (2006) similarly characterized AL as how students use language to perform academic tasks and addressed the unique challenges facing teacher preparation for mainstream teachers in the era of new standards (Bunch, 2013). Part of the challenge in writing a specific definition of AL is that language itself refuses to be

categorized, especially regarding its manifold purposes, which is the central point of Wittgenstein's (2001) theory of the *language game*.

A learner who is trying to make sense of "how things are" in mathematics—rather than expressing a feeling or attitude about mathematics—is what presents the challenge for teachers. In its most simple state, mathematics appropriates the use of otherwise familiar terms (e.g., What is three fourths *of* 16?), and, in perhaps its most complex state, words and relations are represented entirely by variables (e.g., $x = y^2$). Learners are unlikely to gain this specialized language by mere exposure, so teachers must consider how their students can best learn these linguistic conventions.

Despite the recognized importance of teaching students to decode mathematical syntax, comprehend the accompanying vocabulary, and communicate their results effectively, few licensure assessments for mathematics teachers require teacher candidates to demonstrate these abilities. The Performance Assessment for California Teachers (PACT) is the first assessment of teaching to include mastery of AL knowledge by teachers not specializing in teaching ELLs. The decision to include AL teaching proficiency on the PACT followed from a combination of important considerations, including the need to provide a rich education to the diverse California student population, as we discuss further in the following "Background" section. Moreover, to our knowledge, the only other widely used evaluation of teachers that assesses their proficiency of incorporating and/or developing students' AL levels in the classroom is the National Board for Professional Teaching Standards's (NBPTS) Teaching English-as-a-New-Language portfolios for Early and Middle Childhood and Early Adolescence Through Young Adulthood (NBPTS, 2013b, 2014). We point out, however, that these tasks and rubrics are designed to measure superior teaching skills of only those teachers who have chosen to seek NBPTS recognition in the specific domain of teaching ELLs, such as English Language Development Specialists (NBPTS, 2013a). Moreover, we could not find any research documenting the reliability or validity of the AL rubrics for these English-as-a-New-Language portfolios.

With the passage of Senate Bill 2042 in 1998, California's state legislature (Legislative Counsel of California, 1998) mandated that each preparatory institution ensure that its credential candidates meet the Teacher Performance Expectations, a set of standards that aligns with the California Standards for the Teaching Profession set by the California Commission on Teacher Credentialing (CCTC; 2009). Each teacher preparation program in California is required to assess whether its candidates have met the defined California state standards of teaching competencies. In response, the PACT consortium designed and constructed subject-specific performance assessments modeled after the portfolio assessments of the Connecticut State Department of Education, the Interstate New Teacher Assessment and Support Consortium, and the NBPTS (PACT, 2008a) and was approved by the state as one of the possible licensure exams.

The PACT has not only moved from pilot to full implementation in Califor-

nia but has also inspired the birth of a nationwide teaching licensure exam called "edTPA" (edTPA, 2014; Sato, 2014; Stanford Center for Assessment, Learning, and Equity [SCALE], 2015). Thus the full implementation of the PACT in California in a diverse set of California teacher education programs, coupled with the growing interest in a similar assessment across the nation, strongly motivates a current validity study of the PACT.

In this study, we investigated the validity of the internal structure of the PACT with operational data for Elementary Mathematics using multidimensional item response theory (MIRT) models. Such models allowed us to explore the relationships among the PACT content domains (as represented by the scoring rubrics) to determine if and how they are related to one another. In particular, we aimed to determine which and how many distinct constructs the Elementary Mathematics PACT instrument assesses, with a particular interest in how the evolving AL domain behaves in relation to the other domains. We addressed this aim by determining the extent that various MIRT models fit and provide meaningful feedback about teacher candidate performance.

We aim to explore the properties of the PACT overall, but we take a particular focus on the item scores for the AL domain—one of five content domains assessed by the PACT. The general planning–instruction–assessment–reflection model of teacher assessment is at least two decades old (see Collins, 1991), whereas the assessment of AL knowledge and skills in teaching, particularly for teacher licensure, is almost exclusively unique to the PACT, making it a novel domain. Moreover, AL is an evolving domain that is particularly critical to the effective teaching of mathematics to students of all linguistic backgrounds.

As a team of educational researchers, psychometricians, and teacher educators, we recognize the importance of the inclusion of the AL items, rubrics, and exemplars on the PACT for California teacher candidates. The goal of our study is to learn more about the meaning of the AL construct, and thus we focus on a single but important aspect of validity evidence, namely, the internal structure of the Elementary Mathematics PACT, which allows us to answer critical questions about the assessment of AL for elementary mathematics teacher candidates: What does it mean to be AL proficient on the PACT? Which AL tasks are more difficult than others? How, if at all, are AL tasks on the PACT related to those in other content domains? One approach to answering these complex questions is to use measurement models to evaluate the fit between theoretical claims of instrument developers and empirical observations represented by the score data. The best fitting, most informative models can, in turn, provide actionable information on how PACT should move forward with assessing and scoring AL.

Description of the PACT Instrument

Before presenting our study and its results, we provide more background on the PACT instrument and the inclusion of the distinct AL domain. The PACT is designed as an authentic and integrative performance assessment that requires preservice teachers to submit two sets of tasks: the Embedded Signature Assessment (ESA) and the Teaching Event (TE). The ESA is a preparatory program-specific formative assessment and as such is not the focus of this study. Rather, this study focuses on the standardized and summative TE. The TE involves a collection of teaching artifacts for a focused, 1-week teaching and learning segment including lesson plans, video clips of teaching and learning, student work samples, and daily reflections, as well as commentaries responding to a set of task-specific prompts (PACT, 2012b).

The structure of the TE involves tasks, domains, and items, as illustrated in Figure 1, with the tasks corresponding to multifaceted sets of directions and prompts to which candidates respond and submit various materials, whereas the domains and items correspond to how these materials are scored. For instance, for the Context for Learning task, candidates provide descriptions about the instructional context and decisions for their selected learning segment by completing the Context for Learning Form and responding to several prompts about features of their class and how they may affect instructional decisions in a three- to five-page Context Commentary (PACT, 2012b). But these submitted materials are scored with those for the Planning tasks across three items within the Planning domain—items P1 to P3.

Figure 1
Illustration of the structure of the scoring of the PACT Teaching Event

The tasks in the left-hand column consist of questions and prompts to which teacher candidates respond and for which they submit various materials from written commentaries to video segments of teaching. The (scoring) domains and their corresponding items reflect how these materials are assessed by raters.

	(Scoring) Domain				
	Planning (P)	Instruction (I)	Assessment (A)	Reflection (R)	Academic Language (AL)
Task					
Context for Learning	Items 1, 2, 3				Items 11, 12
Planning	Items 1, 2, 3				Items 11, 12
Instruction		Items 4, 5			Items 11, 12
Assessment			Items 6, 7, 8		Items 11, 12
Reflection				Items 9, 10	Items 11, 12

In contrast, the Instruction, Assessment, and Reflection tasks are scored along the domains of their same name, and the AL domain is scored across all tasks. That is, there is no specific AL task to which candidates respond; rather, questions and prompts related to AL are included in all of the tasks. There are 2 to 3 scored items per domain for a total of 12 items.

The items are scored on a 4-point scale ranging from 1 (lowest) to 4 (highest). Centrally trained, subject-specific raters from each local institution assess and score the portfolios of their teacher candidates. Candidates fail if they receive more than one level 1 score for the items within any of the five domains or if they have more than three level 1 scores across all domains (Pecheone & Wei, 2007). Generally, only one rater scores a portfolio, except for double-scoring of candidates who receive a failing or near-failing score from the first rater and for a random sample of candidates to check rater consistency.

Academic Language and the PACT

As previously discussed, the PACT licensure TE is unique in its inclusion of AL scores. The PACT (2012b) consortium generally defines AL as

> the language needed by students to understand and communicate in the academic disciplines. Academic language includes such things as specialized vocabulary, conventional text structures within a field (e.g., essays, lab reports) and other language-related activities typical of classrooms (e.g., expressing disagreement, discussing an issue, asking for clarification). (p. 20)

Moreover, a PACT (2007) scorer training manual emphasizes that the rubrics for these items focus on "academic language both as a medium for learning content and as an independent dimension of content learning" (p. 43), which is in line with having two specific AL scores (Items AL11 and AL12; see Figure 1) that are scored using material from the full teaching portfolio. However, this scoring choice has evolved throughout the life of the PACT.

The decision to include a rubric assessing a candidate's capacity for teaching the AL of the discipline was debated among the PACT developers for some time. Drawing from a growing body of theory and research demonstrating the role of language in disciplinary understanding and expression (Hyland, 2004), the discussions turned not on whether teaching candidates should have, at a minimum, an emerging awareness of AL but rather on if such knowledge and skill could be accurately assessed. Nevertheless, the core group (which included one of the authors of this study) concluded that the AL rubric was needed both to address the content knowledge–specific aspect of the PACT and to push the teacher education community in California toward a new understanding with respect to the discourse of the disciplines.

Moreover, California has a large population of ELs, constituting about 30% of the state's overall student population and even greater proportions at the elementary

grades (for an overview, see Téllez, 2010). Under California Senate Bill 2042, every credential earner—not just those specializing in educating ELs—must be qualified to teach ELs. Additionally, the federal No Child Left Behind legislation demands that ELs meet the same performance standards as their native English–speaking counterparts (Kersaint, Thompson, & Petkova, 2009, pp. 3–4). Charging teachers with the task of instructing both ELs and non-ELs simultaneously, however, could hinder the ELs' academic progress as cultural adaptation and language proficiency gradually develop over time, unless the teachers are able to implement appropriate pedagogical approaches to accommodate these students' language demands and developmental needs (Kersaint et al., 2009). Accordingly, the PACT had to address language teaching in some form. The PACT thus requires that candidates carefully analyze the content-specific language demands of academic tasks while also considering how to make that content accessible to ELs through carefully designed instruction. As Moschkovich suggests, mathematics instructors need to "recognize and strategically support EL students' opportunity to engage with this language complexity" (Moschkovich, 2012, p. 23).

In the mathematics classroom, AL-driven teaching and learning are not merely about vocabulary use and should consider everyday language and experiences as resources (Hakuta, 2013). Thus it is critical that teacher licensure instruments capture the enactment of AL—its use by students, the supports provided by teacher candidates, and the process of exchange between students as they grapple with those demands. The PACT's instructional video, tasks, and rubrics were expected to provide "enactment" (as opposed to mere planning or reflecting) evidence for teacher candidates' placement on the AL construct.

In the first year of the pilot, 2002–2003, the PACT involved a rubric focused solely on ELs for each of the four content domains—Planning, Instruction, Assessment, and Reflection. However, initial feedback and early pilot data suggested there was insufficient evidence to support so many rubrics (PACT, 2006). Moreover, teacher candidates expressed frustration with focusing on only ELs when they had non-ELs who also had difficulties with formal AL (PACT, 2008a). The core designers thus revised the structure of the PACT, adding AL-specific rubrics that draw on evidence from each of the PACT tasks, which reduces the number of AL rubrics but still emphasizes the need for accommodating AL proficiencies of their students through all stages of the teaching process from planning to reflection (PACT, 2008a). In general, however, the PACT developers have struggled to create rubrics that distinguish between candidates who have mastered advanced understanding and teaching of AL and those who hold only a thin understanding of the concept. The variations in the rubrics over the years are evidence of this challenge. For instance, the AL11 item has shifted focus from candidates demonstrating that they can accommodate any AL proficiency in the 2008–2009 PACT, to accommodating only ELs in the 2009–2010 academic year, and then back to students at different academic language proficiencies in 2012–2013 (PACT, 2008b, 2009, 2012a, 2012b).

We also suggest that the PACT AL rubrics are written in such a way that candidates are drawn to paying close attention to teaching academic vocabulary while ignoring other features of AL (e.g., analyzing text types or designing lessons to explore genre-specific meanings and uses). With some ease, a candidate can earn a score of 2 on the PACT's 4-point scale—a score that is just good enough to pass but not at the high end of the scale (i.e., scores of 3 or 4). Indeed, in our data sample (described in the following section), about 59% and 53% of candidates received scores of 2 on AL11 and AL12, respectively, compared to 20% to 47% of candidates receiving a scores of 2 on all the other items. For the other items, there were generally at least 50% of candidates earning scores of 3 or 4, whereas only 27% and 42% of candidates earned these higher scores on items AL11 and AL12, respectively. These low scores may reflect uncertainty over the demands of the AL items among both candidates and scorers.

Previous Validity and Reliability Studies

To place our study in the context of other validity and reliability studies on the PACT, we briefly review previous studies. Pecheone and Wei (2007) conducted the most extensive prior PACT validity study, in which they investigated several strands of evidence, including content validity, bias and fairness, construct validity, criterion-related concurrent validity, score consistency, and reliability. They used pilot score data from 2003–2004 for 625 submitted portfolios for various subject-specific TEs, including the Elementary Mathematics TE. Their study generally yielded positive results, prompting them to recommend the use of the PACT operationally. In particular, their fairness/bias review, using only the 46% of their sample that had matched score and demographic data, found no significant differences between scores by candidates' race/ethnicity, percentage of ELLs, grade level taught, students' academic achievement level, or months of previous paid teaching experience. They did, however, find some meaningful differences: Women significantly outscored men on average, and candidates teaching in high-socioeconomic, suburban schools outscored those teaching in low-socioeconomic, urban or inner-city schools.

Similar to our primary aim of seeking to determine the meaningful, distinct constructs assessed by the PACT, Pecheone and Wei (2007) investigated construct validity evidence for the Elementary Mathematics TE with exploratory factor analysis. They found evidence for two distinct factors—one for Planning, Instruction, and Academic Language and another for Assessment and Reflection—indicating that the test was tapping into distinct constructs of teaching, but not as many as those used in scoring the test (see Figure 1).

Bunch, Aguirre, and Téllez (2009) conducted a small, in-depth qualitative study to examine AL exclusively. They analyzed the specific texts of elementary mathematics candidates' PACT TEs and found that only two of eight candidates explored AL in any depth beyond introducing vocabulary germane to the mathematics lesson.

Duckor, Castellano, Téllez, Wihardini, and Wilson (2014) analyzed the internal structure of the Elementary Literacy TE with a large sample (n = 1,711) of teacher candidates from several California teacher preparation programs. They found that item scores were well explained by a unidimensional, polytomous IRT model. They also explored relationships among the content domains with MIRT models, finding evidence of a three-dimensional model with separate dimensions for Planning and Instruction and a combined dimension of Assessment, Reflection, and Academic Language, or "Meta-Reflection."

Other studies have explored specific aspects of the validity of the PACT. Sandholtz and Shea (2012) explored the relationship between supervisors' predictions and candidates' performance on the PACT. The results indicated that university supervisors' predictions were not closely associated with PACT scores, particularly for high and low performers. This finding may suggest that PACT lacks concurrent validity or consistent interpretations about teacher readiness as supervisors' predictions. However, the authors posited an alternative explanation: that the university supervisors and PACT scorers are drawing from different sources of information over different time points in making their evaluations and thus may offer useful distinct information about aspects of candidates' readiness to teach. Their research suggests that the use of multiple measures should be considered in evaluations of candidates' readiness to teach.

Okhremtchouk et al. (2009) found that candidates viewed the PACT as helpful in improving their instructional practice. This study may offer a measure of face validity for the PACT, demonstrating that candidates believed the PACT helped them to develop their teaching, but it did not link such perceptions to candidate performance on the PACT.

Darling-Hammond, Newton, and Wei (2010) argued for positive triangulation of the PACT data with several other measures of student teacher learning to augment information needed to make useful and effective decisions for improvement of a teacher education program. These researchers also conducted a predictive validity study relating the preservice teachers' PACT scores to their later teaching effectiveness in ELA and mathematics at Grades 3–8, as measured by standardized test scores (Darling-Hammond, Newton, & Wei, 2013). They found significantly positive relationships between PACT subscores and the students' California standardized test scores to varying degrees.[1] The assessment domain score was found to be a strong predictor of effective teaching on both ELA and mathematics, whereas the score on the planning domain was more predictive for ELA only.

Although previous studies of the PACT have looked at issues related to validity, concerns about reliability (e.g., drift, "halo" effects) have been less well documented. Porter (2010) demonstrated that interrater reliabilities—summarizing the consistency of scores across different raters—for the PACT were poor to moderate for local score data.

Our study uses formal measurement models to investigate the internal structure

of the PACT using a large sample of operational test scores for candidates from several teacher preparation programs. Our approach follows the professional *Testing Standards*, which define "internal structure validity evidence" as referring to "the degree to which the relationships among test items and test components conform to the construct on which the proposed [instrument] score interpretations are based" (American Educational Research Association, American Psychology Association, & National Council on Measurement in Education, 1999, p. 13). That is, for instance, if a score is associated with a word such as "planning skill," then it should show evidence of planning skills. Note that our focus is on this particular type of validity evidence, which allows us to directly address our research aims; however, other sources of validity evidence are also important to collect. Pecheone and Wei's (2007) study, for instance, investigated several aspects of validity for all the PACT subject TEs, but each of these should be periodically revisited as the PACT evolves over time. Moreover, we can look to validity studies for other performance assessments as examples, such as Wilson, Hallam, Pecheone, and Moss's (2014) rigorous external validity study of the Connecticut performance-based teacher assessment.

We investigate the claims by PACT test designers by examining the extent that empirically observed relationships among the PACT scores for items within and across content domains (Planning, Instruction, Assessment, Reflection, and Academic Language) represent those intended by the scoring rubrics, by the test's scoring structure as illustrated in Figure 1, and as documented in the descriptive materials for the PACT licensure exam. Although our study is similar to the Duckor et al. (2014) study on Elementary Literacy, it differs in our more focused analysis of AL in Elementary Mathematics teacher credentialing generally and the behavior of this domain in the PACT instrument specifically. Our study also differs from the Pecheone and Wei (2007) study of the structure of the PACT, as they used pilot data and exploratory factor analysis, whereas we use operational data and MIRT to determine which teaching-readiness constructs are meaningfully assessed by the PACT. A MIRT measurement modeling approach is advantageous in that it more appropriately models the (ordered) categorical nature of the item data (i.e., the 1- to 4-point structure), and it allows us to determine how measurement qualities such as item and person fit statistics and differential item functioning (DIF) are affecting the PACT score results.

Methodology and Methods

Data Sample

In this study, we solicited participating public institutions that administer the PACT licensure exam. We obtained Elementary Mathematics TE data from five teacher preparatory programs at different University of California institutions. The data set included item-level scores for all 505 teacher candidates who completed

the Elementary Mathematics TE in the 2008–2009 and 2009–2010 academic years. Unfortunately, no examinee-, rater-, or institutional-level descriptive variables were provided in accordance with the scope of consent obtained for this study; we discuss the limitations of these data constraints in the conclusion to the article. 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 thus share these outcome goals. Although programs vary in size and geographical location, the data sample is consistent with the population of public programs across the state.

Table 1 provides summary statistics by item for each administration year and overall. The mean item scores range from 2.16 to 2.93, with the Planning items as the easiest and AL Item 11 as the most difficult at both time points. For all of the items, the majority of the scores are 2 or 3. Looking across the 12 items, approximately 1% to 12% of the item scores are 1, and 4% to 21% are 4. Generally, there are complete data for all items, with the one exception of 66 missing scores for the eighth Assessment item (A8), which mostly occurred for examinees at a single campus, and only one or two missing scores for other items.

We used qualitative and quantitative data checks to ensure that the wording and structure of the instrument itself were constant over the two test administrations. If the items function the same substantively and statistically across the two time points, then we can use the full sample size (n = 505) when we fit each model, which gives us more statistical power to test the relationships among the items. Through a DIF procedure that involved fitting the unidimensional model

Table 1

Summary Statistics by Item

		2008–2009			2009–2010			Overall		
Domain	Item	*N*	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Planning	P1	102	2.98	0.69	402	2.92	0.64	504	2.93	0.65
	P2	102	2.83	0.81	403	2.89	0.74	505	2.88	0.76
	P3	102	2.75	0.74	403	2.78	0.68	505	2.77	0.69
Instruction	I4	102	2.52	0.67	402	2.67	0.71	504	2.64	0.70
	I5	102	2.48	0.82	402	2.53	0.81	504	2.52	0.81
Assessment	A6	102	2.76	0.86	403	2.69	0.80	505	2.71	0.81
	A7	102	2.32	0.83	403	2.43	0.75	505	2.41	0.77
	A8	83	2.23	0.75	356	2.56	0.83	439	2.50	0.82
Reflection	R9	102	2.56	0.77	402	2.67	0.72	504	2.65	0.73
	R10	102	2.53	0.83	403	2.56	0.73	505	2.55	0.75
Academic	AL11	102	2.21	0.65	403	2.14	0.72	505	2.16	0.70
Language	AL12	102	2.30	0.66	402	2.42	0.64	504	2.40	0.65

separately by administration time,[2] we found that Item A8 behaved differently over time. A qualitative review of the items revealed that Item AL11 substantively shifted focus from candidates describing language demands for students with any student language development impediment to *only* ELs; thus, although the item difficulty did not change significantly, the item itself changed. Accordingly, we combined the data sets and treated A8 and AL11 as two separate items by test administration period.[3]

Data Analysis

As illustrated in Figure 1, the structure of the Elementary Mathematics TE suggests that the instrument evaluates candidates on multiple constructs, each contributing to a decision about readiness to teach in the California classroom. Owing to the importance of preparing teachers to support different language levels and proficiencies in California, we were particularly interested in how the academic language domain behaves in relation to the other domains represented by the PACT instrument (Pecheone & Wei, 2007). MIRT analyses can reveal important information about how the AL items are functioning and how they are best interpreted in relation to the instrument's proposed uses. A multidimensional analysis of the internal structure of the PACT can also offer clues about how to either restructure the PACT instrument to better capture the AL dimension or refocus rater training so that the scoring of AL items is more reliable.

Specifically, we used the multidimensional version of the partial credit model (PCM) for polytomous items. PCM is within the Rasch family of IRT models and thus has the advantage that it can reflect the differences in the difficulty among test items and present the distribution of the test takers on the same scale. In the multidimensional PCM, person *n*'s latent ability estimate in dimension *d* (θ_{nd}) is calculated from the probability of success of answering an item *i* in *X*=*x* response category (*x* = 0,1,…,*m*), which is a function of the difference between the person *n* location and the item *i* location. Specifically, the model is as follows:

$$P(X_{ni} = x|\boldsymbol{\theta}_{nd}) = \frac{\exp\left(\sum_{j=0}^{x}((\sum_{d=1}^{D}\theta_{nd}) - \delta_{ij})\right)}{\sum_{k=0}^{m}\exp(\sum_{j=0}^{k}((\sum_{d=1}^{D}\theta_{nd})) - \delta_{ij}))}$$

Here *d* indicates a specific latent dimension (i.e., *d* = 1,…,*D*); θ_{nd} represents person *n*'s latent ability parameter on dimension/construct *d*; and δ_{ij} is the item-step difficulty parameter for item *i* at category *j* (i.e., *j* = 0,…,*k*,…*m*; Wilson, 2005; Wright & Masters, 1982).

We first fit the unidimensional model as a point of reference for the MIRT models. Subsequently, we assessed the fit and utility of the task-based model, the domain-based model, and other models driven by empirical findings and theoretical hypotheses. We define and discuss each of these in turn in the following subsections.

To fit all models, we used the psychometric computer program ConQuest (Adams, Wu, & Wilson, 2012).

Results

The Unidimensional Model

The unidimensional model provides a single teaching-readiness ability estimate for each teacher candidate. This is the model that is most suitable for the actual usage to which the PACT scores are put: providing a single criterion of teacher readiness. However, the Elementary Mathematics TE is scored on five different tasks and domains, as shown in Figure 1. Examination of the weighted mean square item fit statistics revealed good model fit (Adams & Khoo, 1996; Wilson, 2005). However, this model does not provide information on teacher candidate "skills" and "proficiencies" on different aspects of the content embodied in the TE.

The Task-Based Model

We first assessed the structure of the Elementary Mathematics TE with a multidimensional model that matches the TE's scoring structure illustrated in Figure 1. In this task-based factor structure, as shown in Figure 2a, the model has four dimensions corresponding to the five tasks (note that the first two tasks—Context for Learning and Planning—both correspond to the Planning domain). For this model, the Planning, Instruction, Assessment, and Reflection items each mapped onto different dimensions, but the Academic Language items loaded onto all of the dimensions. Although we expected that this model would fit well as it follows the intended structure of the TE, we found it resulted in relatively poor model fit.

To assess global model fit, we compared the Akaike information criterion (AIC) of this model to the unidimensional PCM, with smaller values indicating better fit. The AIC of the task-based model was 10,679 versus an AIC of 10,416 for the unidimensional model, indicating that the task-based multidimensional model fit worse than the unidimensional model. Moreover, the individual item (weighted mean square) fit statistics for the AL items were outside of the usual acceptable bounds (0.75–1.3; Adams & Khoo, 1996). Specifically, the AL items had high item fit statistics (approximately between 1.5 and 1.9), indicating that these items have 50% to 90% more variation in their scores than predicted by the model or that the model underfits the variation in these items. This result demonstrated that, although the items were designed according to Figure 1, the resulting data were not consistent with this test structure.

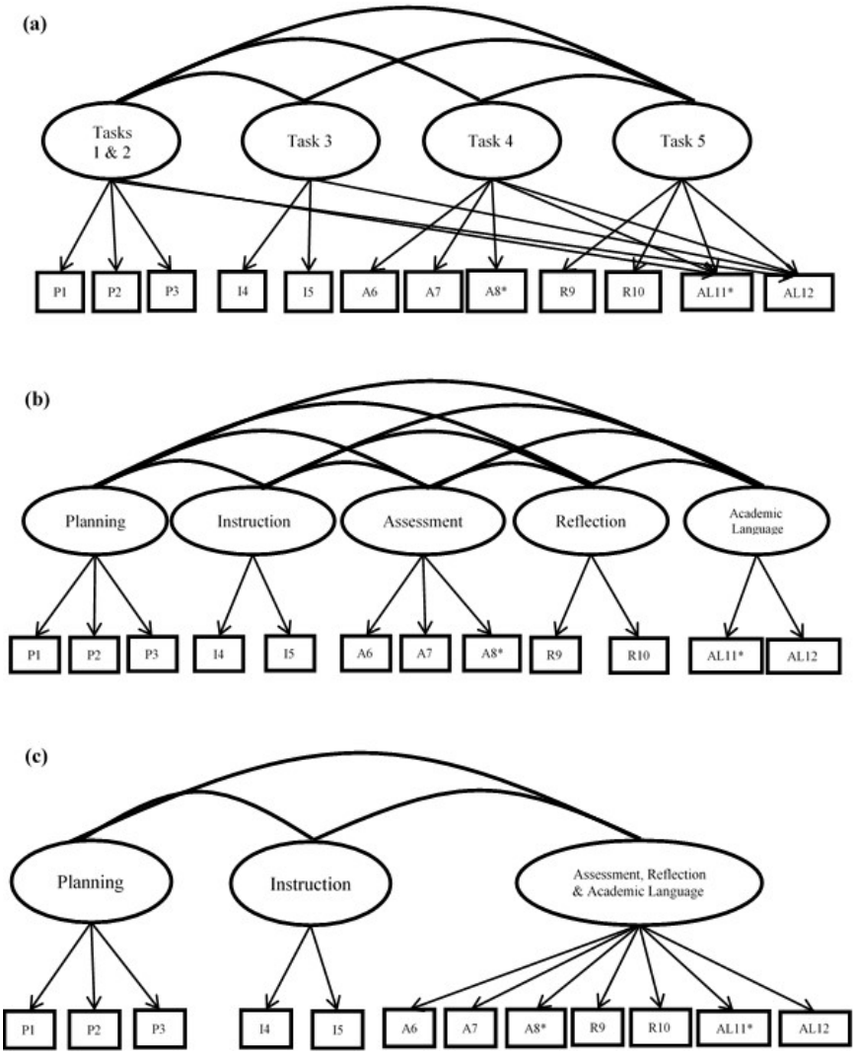
The Domain-Based Model

The misfit of the AL items for the task-based multidimensional model suggested that raters may not have used all of the materials across all the tasks to score

Figure 2

Illustration of the 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 (P), the second by Instruction items (I), and the third by Assessment (A), Reflection (R), and Academic Language (AL) items. Note: Assessment Item 8 (A8) and Academic Language Item 11 (AL11) are treated as two separate items by administration year.



these items as instructed. Or, as a relatively new or conceptually difficult construct, the AL items may make additional demands on teacher candidates, raters, and/or preparation programs. The AL items may thus represent their own dimension. To test this hypothesis, we fit a five-dimensional (5D) "domain-based" model with items for each domain mapping onto its own dimension, as illustrated in Figure 2b. Unlike the task-based model, for this model, each item only contributes to one dimension. This model resulted in good model fit compared to both the task-based model and the unidimensional model (e.g., AIC_{5D} = 10,143 vs. AIC_{uni} = 10,416). In addition, the individual item fit statistics were all within acceptable bounds.

Other Domain-Based Models

We further hypothesized other possible domain-based models with fewer than five dimensions that might better reveal substantively meaningful dimensions. Primarily, we hypothesized that Assessment, Academic Language, and Reflection composed a single construct of teacher readiness as assessed by the PACT Elementary Mathematics TE. The theoretical rationale for treating Assessment, Academic Language, and Reflection as distinct domains of teaching practice is well documented and supported by experts. For novices writing about their own teaching practices and beliefs about, for example, the role of AL in teaching math, the literature is less robust and definitive. Discourse analyses show that teachers struggle "in the moment" with managing the social and interpretive process of student learning (Barwell, 2005). Compounding the challenges inherent in teaching math discursively is the demand for assessing EL in the heterogeneous classroom in ways that are consistent and meaningful (Moschkovich, 2007, 2013). The fact that the teacher candidate is expected to reflect upon the AL, assessment, and theory-laden components of the TE after the fact leads to further problems related to metacognition. Skills involving self-regulation, goal setting, and even the ability to understand, control, and manipulate one's cognitive processes are fundamental to success (Meichenbaum, 1985; Olafson, Schraw, & Vanderveldt, 2010; Schraw, 1998). As it is currently structured, the PACT demands that the novice teacher candidate write a persuasive rationale, which we call a meta-reflection, for intersecting and often confusing elements of practice.[4]

In addition to our theoretical rationale for hypothesizing that AL can be combined with Assessment and Reflection as one Meta-Reflection domain, we found empirical evidence supporting this hypothesis through an analysis of the correlations estimated for the 5D domain-based model. The disattenuated correlations estimated for the 5D domain-based model are given in Table 2 (below the diagonal), along with the correlations among the domain scores (not corrected for measurement error). The disattenuated correlations range from .75 to .92. The strongest pair-wise correlations were among the Assessment, Reflection, and AL domains (ranging from .84 to .92), suggesting they may be collapsed into one dimension with minimal loss

of information. Teacher candidates who score highly on one of these dimensions tend to score highly on the other two, and vice versa. The Planning and Instruction dimensions are each correlated about .75 with each of the other dimensions and each other and so provide somewhat more distinct information about an aspect of readiness to teach.

To test our hypothesis, we fit a 3D modified-domain-based model with Planning and Instruction as their own dimensions and Assessment, Reflection, and AL as a single Meta-Reflecting dimension (see Figure 2c). This model has good item fit, but with a higher AIC value, meaning it does not exhibit as good model fit as the domain-based 5D model (AIC_{3D} = 10,185 vs. AIC_{5D} = 10,143). However, the correlations among the dimensions in the 3D model support this model as providing more distinct information on candidate ability than the 5D model. For the modified-domain-based 3D model, the Meta-Reflecting dimension is correlated .796 with the Planning dimension and .790 with Instruction, whereas the Planning and Instruction dimensions are correlated .751. All of these disattenuated correlations are lower than the three among Assessment, Reflection, and AL in the 5D model. Thus the 3D model's dimensions are more distinct than the five dimensions in the 5D domain-based model. We also found that each dimension has as high or higher reliability estimates than those for the 5D model.

This modified-domain-based model also fits better than any other hypothesized modified-domain-based models we fit. For instance, we considered Pecheone and Wei's (2007) 2D model with Planning, Instruction, and AL domains constituting one dimension and Assessment and Reflection the second dimension. This 2D model did not fit as well as our 3D model (AIC_{2D} = 10,329 vs. AIC_{3D} = 10,185). Given that Pecheone and Wei used pilot data, the number of items per domain and some item wording have changed since then, and they used a different modeling approach (factor analysis vs. MIRT), it is not surprising that we found evidence of a different internal structure for the Elementary Mathematics TE.

We also tried fitting a 2D model with AL items mapping to their own dimen-

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

	Mean domain scores				
Disattenuated Correlations	Planning	Instruction	Assessment	Reflection	Academic Language
Planning		.60	.59	.59	.59
Instruction	.75		.58	.58	.57
Assessment	.76	.75		.71	.62
Reflection	.75	.77	.92		.65
Academic Lang.	.76	.77	.84	.89	

sion and all other items mapping to the second dimension. This model's AIC was also greater at 10,392. Accordingly, the 3D modified-domain-based model with AL, Assessment, and Reflection collapsed as one dimension better reflected the relationship AL had with the other domains. It also provided evidence that the TE is assessing different aspects of the teaching process, but not necessarily as intended by the PACT instrument developers.

Given the fit and utility of the 3D model, we further explored how it characterized the internal structure of the Elementary Mathematics TE using a Wright map (Wilson, 2005) after applying delta-dimensional alignment to place items from all three dimensions on the same scale (Schwartz, 2012).[5] This Wright map, shown in Figure 3, shows the distributions of the teacher candidate proficiency estimates (left) for each dimension on the same logit scale as the Thurstonian thresholds for the item-step difficulties (right). These item thresholds are denoted as *i.k* for item *i* at score level *k* and are defined as the location on the latent ability scale at which candidates have a 50% chance of scoring at or above level *k* for item *i* (Wu, Adams, Wilson, & Haldane, 2007). Figure 3 clearly shows that teacher candidates of all ability levels on the Planning dimension (first column) had at least a 50% chance of obtaining a 2 or higher on the Planning items, whereas this is not the case for the other dimensions. Comparing the item-step difficulties across the three dimensions, it appears that getting a score 3 or 4 on AL items for a teacher candidate was more difficult than it was on the other teaching domains. We also note that although we treated AL11 as separate items by administration year, the item thresholds are very similar for AL11a and AL11b, particularly for the third and fourth thresholds. Accordingly, although these items differ substantively, they are functioning similarly for teacher candidates in the 2 years, which may indicate that candidates and raters responded to and scored them in the same way despite the change in focus from all students to only ELL students.

Discussion

Our study investigated the internal structure (i.e., the dimensionality) of the Elementary Mathematics TE for Tier I licensure in California. Using MIRT models, we found that Planning and Instruction are meaningfully distinct dimensions that correspond with the content validity arguments advanced by the PACT developers (Pecheone & Wei, 2007). However, we also found that Assessment, Reflection, and AL domains in the Elementary Mathematics TE are tapping into very similar "skills" and "proficiencies," which may make it difficult to discern the meanings of scores on these tasks. Our findings with regard to the AL construct indicate that score interpretation and use of subscores should proceed with caution.

One strategy for addressing the problem of internal structure validity is to simply embrace the factor or dimensional "solution" provided by the model fit statistics. Accordingly, one treats the difficulty with validly interpreting AL score

data (rather than the potential constraints from the items or scoring design itself) as the problem. The edTPA collapses the construct of AL into subtasks within the Planning, Instruction, and Assessment teaching domains, perhaps to avoid these dimensionality issues. Interestingly, this supposed solution results in score data from the edTPA licensure exam that emphasize teacher candidates' knowledge, skills, and use of AL in only the Planning and Assessment domains (SCALE, 2013).

The edTPA focuses on how teacher candidates are supporting their students' language demands by identifying the use of vocabulary on the language function

Figure 3

A Wright map of the three-dimensional domain-based model using elementary mathematics data

Each "X" represents 4.2 teacher candidates.

	Dimension 1: Planning		Dimension 2: Instruction		Dimension 3: Meta-Reflection		
	Person Ability	Item Score Threshold	Person Ability	Item Score Threshold	Person Ability	Item Score Threshold	
5						AL12.4	
	X		X		X	AL11a.4,AL11b.4	Threshold for Score Level 4 (ability level at which examinee has 50% chance of obtaining a score of 4 or higher)
			X		X		
	X		XX		X		
4	XX		X			A7.4	
	X		X	I4.4		A8a.4	
	X	P3.4	XX	I5.4	X	R9.4,R10.4	
3	XXX		XX		XXXX	A8b.4	
	XXX	P1.4	XXX		XXXX		
	XXXX	P2.4	XXXX		XXXXX	A6.4	
	XXXX		XXXX		XXX		
2	XXXXXX		XXXXX		XXXXXXX	AL11a.3,AL11b.3	
	XXXX		XXXXXXXX		XXXXXX		
	XXXXXXX		XXXXXXXX		XXXXXX	A8a.3	
1	XXXXXXXXXX		XXXXXXX		XXXXXX		Threshold for Score Level 3 (ability level at which examinee has 50% chance of obtaining a score of 3 or higher)
	XXXXXXXX		XXXXXX		XXXXXXX	A7.3, AL12.3	
	XXXXXXXX		XXXXX	I5.3	XXXXXXXX	R10.3	
0	XXXXX		XXXXXXX		XXXXXXX	A6.3, A8b.3	
	XXXXXX		XXXXXXX	I4.3	XXX	R9.3	
	XXXXXX		XXXXX		XXXXXX		
	XXXXX	P2.3,P3.3	XXXXX		XXXXXXXX		
-1	XXXXXX		XXXXXX		XXXXXXX		
	XX		XXX		XXXXX		
	XXXXXX	P1.3	XXXXXXXX		XXXXXX		
-2	XXXX		XXXXX		XXXXXX		
	XXXX		XXXXX		XXX	AL11b.2	
	XXX		XX		XX		
	XXX		XXX		X	A7.2	Threshold for Score Level 2 (ability level at which examinee has 50% chance of obtaining a score of 2 or higher)
-3	X		X		X	A8a.2, A8b.2	
	XX		X	I5.2		AL11a.2	
	X					AL12.2	
-4			X		X	R10.2	
				I4.2	X	A6.2, R9.2	
		P3.2					
-5		P1.2,P2.2					

in the Planning domain and how they are analyzing their students' use of language for content understanding in the Assessment domain. However, the lack of instrumentation targeted on observing actual enactments of AL skills, capacities, and proficiencies of these preservice teachers in the Instructional domain is a blind spot. The *enactment* of academic language–driven instruction is deemphasized. Moreover, any licensure exam that inadequately addresses (in part, by inadequately observing) the importance of the AL construct in mathematics instruction seems to contradict both the robust findings in the research literature and the new policy direction that focuses on speaking, listening, and other modalities of productive language instruction under the Common Core framework (Hakuta, 2013; Heritage et al., 2015; Moschkovich, 2012).

The results of the Elementary Mathematics TE in this study are similar to those reported in a previous validation study on the English Language Arts TE (Duckor et al., 2014). Thus, although our study is limited by its voluntary sample of California teacher preparation programs and its sample size did not allow for split-data analysis, its findings are consistent with a separate study of a different data sample for a different PACT TE. In both studies, the implications for policy and practice in the context of the PACT licensure exam are varied. Data-driven state policy makers and teacher educators are increasingly compelled to use these results 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, cooperating teachers, program administrators) with respect to the global and subscore data provided by the PACT and other teacher performance assessments. Still others may be tempted to drop the focus on AL in teacher performance-based assessments because it is a conceptually difficult construct to assess. However, we assert that dimensionality studies like ours can justify the meaning of score results. We also advocate the collection of multiple sources of evidence both replicating our own study with other PACT data and even by types of teacher candidates, which was not possible in this case, with the lack of teacher covariates, as well as by collecting further types of validity evidence, such as predictive and consequential validity.

Our research on PACT data suggests that although the AL domain is difficult to distinguish, perhaps because it is closely related to other teaching competencies, its importance to the field is clear. Elementary school learners face increasing pressure to master challenging mathematical concepts, especially those that are related to success in algebra. We know that students who do not master algebra before the ninth grade tend not to take the classes required to attend college. Recent studies have demonstrated that ELLs are particularly at risk of missing key courses during their high school experience (Mosqueda, 2010). Without early success in mathematics, ELLs are effectively pushed out of college consideration as a consequence of course-taking patterns. If elementary teachers cannot make mathematics content accessible to their students, and to their ELLs in particular, the consequences will be far-reaching.

No longer can prospective elementary teachers view mathematics as language free. They must develop the pedagogical skills that link language and mathematics in ways that deepen their students' conceptual understandings, even if experts cannot agree on a single definition of AL (e.g., Snow, 2010). Wittgenstein (1970) wryly wrote that "to understand sums in elementary school, the children would have to be important philosophers. Failing that, they need practice" (p. 122). Wittgenstein is pointing out that even the simplest of mathematical operations lead us to challenging questions that require a comprehensive symbol system to understand, but it is this philosophical link between language and mathematics that contemporary teachers must consider. Preservice teacher evaluation systems in California, such as the edTPA and PACT, must be designed to detect whether teacher candidates possess the skills, knowledge, and dispositions toward practice to help their K–12 students master challenging content.

Conclusions

Teacher licensure exams, such as the PACT and edTPA, as gatekeepers for the teaching profession are designed to ensure that teaching candidates possess the baseline skills necessary to help their K–12 students master challenging subject content. In California, it would be a step backward, given the student population's needs, to shy away from the growing body of research on the intersection of mathematics content and AL demands embedded in the new standards. Our findings on the unintended behavior of the AL items (and their noisy interaction with the Assessment and Reflection domains) warrant further investigation but not an abandonment of the construct itself. Based on our findings, it is likely that the AL instrumentation (i.e., tasks, scoring rubrics, rater training, and/or exam protocols) requires better alignment to the PACT's intended structure. But we also need more data on effects that may be related to examinee, rater, or institutional factors in the AL domain. The PACT consortium could provide a platform for principled scientific investigation of AL at scale, now that we have learned new lessons in California.

This study represents a step in the direction of broadening standards-based validity investigations of the PACT or any teacher performance-based instrument, specifically with respect to particular interpretations about elementary teachers' preparation in the academic language domain. Despite our current, albeit limited, understanding of how to best evaluate AL in novice teachers' work, it remains a critical piece of the puzzle of what it means to be a teaching professional in schools with a commitment to equity and excellence. Beginning teachers can benefit from future work (from educational researchers, measurement specialists, and, most importantly, teacher educators) on how to best assess academic language in their emerging K–12 classroom practice. Policy makers remind us that the success of today's educational reforms may in fact depend on several important shifts in beginning and veteran educators' perspectives on language and language learning in

the early K–8 mathematics classroom. The PACT, with its emphasis on AL across domains of teaching practice, does that to a degree, but, as this study suggests, it can still do more.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305B110017 to the University of California, Berkeley. The opinions expressed are those of the authors and do not represent the views of the Institute of Education Sciences or the U.S. Department of Education.

Editorial Note

The peer review and acceptance of this manuscript was conducted entirely by Reyes Quezada, associate editor of *Teacher Education Quarterly*. At no point was Kip Téllez, editor of the journal, a part of that review process or contacted during the process.

Notes

[1] We note that similar studies with scores for the newly adopted Smarter Balanced Assessment (the current standardized ELA and Mathematics K–12 test in California) may yield different results, which underscores that validation is an ongoing effort that proceeds as new information becomes available.

[2] We used Wright and Masters's (1982) DIF procedure to check if the items behaved the same in terms of difficulty level at both time points. This procedure, which involves plotting the item difficulties at Time Point 1 against the item difficulties at Time Point 2 and computing 95% confidence intervals for the mean difficulty over the two time points, resulted in Item A8 being flagged as having differential difficultly over time (because its point fell outside of the confidence bounds).

[3] We denote A8a and AL11a for responses to these items in the first test administration year and A8b and AL11b for the second administration year.

[4] This finding is not entirely surprising given the PACT's roots in NBPTS and a particular vision for teacher assessment. See, for example, Shulman (1987): "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. . . . We will emphasize teaching as comprehension and reasoning, as transformation and reflection" (p. 13).

[5] Because the means of item and item-step difficulties are set to zero on every dimension in identifying the item parameters of the multidimensional models, the magnitudes of the item difficulties are not comparable across dimensions. The delta-dimensional alignment method provides a means for placing all of the item difficulties on the same scale (for more details, see Schwartz, 2012).

References

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

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

American Educational Research Association, American Psychology Association, & Na-

tional Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: Author.

Bailey, A. L., & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K–12 education: A design document* (CSE Technical Report No. 611). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Barwell, R. (2005). Ambiguity in the mathematics classroom. *Language and Education, 19*(2), 117–125.

Bunch, G. C. (2006). "Academic English" in the 7th grade: Broadening the lens, expanding access. *Journal of English for Academic Purposes, 5*(4), 284–301.

Bunch, G. (2013). Pedagogical language knowledge: Preparing mainstream teachers for English learners in the new standards era. *Review of Research in Education, 37*, 298–341.

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. (2009). *California standards for the teaching profession*. Retrieved from http://www.ctc.ca.gov/educator-prep/standards/CSTP-2009.pdf

Chamot, A. U., & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.

Collins, A. (1991). Portfolios for biology teacher assessment. *Journal of Personnel Evaluation in Education, 5*(2), 147–168.

Common Core State Standards Initiative. (2010a). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/

Common Core State Standards Initiative. (2010b). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/

Danesi, M. (2003). Conceptual metaphor theory and the teaching of mathematics: Findings of a pilot project. *Semiotica, 145*, 71–83.

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.

Darling-Hammond, L., Newton, S., & Wei, R. C. (2013). Developing and assessing beginning teacher effectiveness: The potential of performance assessments. *Educational Assessment, Evaluation, and Accountability, 25*(3), 179–204.

Duckor, B., Castellano, K. E., Téllez, K., Wihardini, D., & Wilson, M. (2014). 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, 65*, 402–420. doi:10.1177/0022487114542517

edTPA. (2014). *Current status of the project*. Retrieved from http://edtpa.aacte.org/about-edtpa#ProjectStatus-2

Hakuta, K. (2013). *Assessment of content and language in light of the new standards: Challenges and opportunities for English language learners*. Retrieved from http://www.gordoncommission.org/rsc/pdf/hakuta_assessment_content_language_standards_challenges_opportunities.pdf

Heritage, M., Walqui, A., & Linquanti, R. (2015). *English language learners and the new standards: Developing language, content knowledge, and analytical practices in the*

classroom. Boston, MA: Harvard Education Press.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor, MI: University of Michigan Press.

Kersaint, G., Thompson, D. R., & Petkova, M. (2009). *Teaching mathematics to English language learners*. New York, NY: Taylor and Francis.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and language for English language learners in relation to Next Generation Science Standards and with implications for Common Core State Standards for English language arts and mathematics. *Educational Researcher, 42*(4), 223–233.

Legislative Counsel of California. (1998). *Teacher credentialing* (Senate Bill 2042). Retrieved from http://www.leginfo.ca.gov/pub/97-98/bill/sen/sb_2001-2050/sb_2042_bill_19980918_chaptered.html

MacGregor, M., & Price, E. (1999). An exploration of aspects of language proficiency and algebra learning. *Journal for Research in Mathematics Education, 30*, 449–467.

Meichenbaum, D. (1985). Teaching thinking: A cognitive-behavioral perspective. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Vol. 2. Research and open questions* (pp. 407–426). Hillsdale, NJ: Lawrence Erlbaum Associates.

Morgan, C. (1998). *Writing mathematically: The discourse of "investigation"* (Vol. 9). Bristol, PA: Falmer Press/Taylor & Francis.

Moschkovich, J. N. (2007). Beyond words to mathematical content: Assessing English learners in the mathematics classroom. In A. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 345–352). New York, NY: Cambridge University Press.

Moschkovich, J. (2012). *Mathematics, the Common Core, and language: Recommendations for mathematics instruction for ELs aligned with the Common Core*. Paper presented at the Understanding Language Conference, Stanford, CA. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/02-JMoschkovich%20Math%20FINAL_bound%20with%20appendix.pdf

Moschkovich, J. (2013). Principles and guidelines for equitable mathematics teaching practices and materials for English language learners. *Journal of Urban Mathematics Education, 6*(1), 45–57.

Mosqueda, E. (2010). Compounding inequalities: English proficiency and tracking and their relation to mathematics performance among Latina/o secondary school youth. *Journal of Urban Mathematics Education, 3*(1), 57–81.

National Board for Professional Teaching Standards. (2013a). *Early and middle childhood: English as a new language: Assessment at a glance*. Retrieved from http://www.nbpts.org/sites/default/files/documents/certificates/Aaag/EMC_ENL_AssessAtaGlance_05.22.13_Final.pdf

National Board for Professional Teaching Standards. (2013b). *Early and middle childhood: English as a new language: Portfolio instructions*. Retrieved from http://www.nbpts.org/sites/default/files/documents/certificates/PF-Instructions/EMC_ENL_Portfolio_Instructions_072313.pdf

National Board for Professional Teaching Standards. (2014). *Early adolescence through young adulthood: English as a new language: Portfolio instructions*. Retrieved from http://boardcertifiedteachers.org/sites/default/files/ENL_EAYA_Portfolio_Instructions_FINAL.pdf

Okhremtchouk, I., Seiki, S., Gilliland, B., Atch, C., Wallace, M., & Kato, A. (2009). Voices of pre-service teachers: Perspectives on the Performance Assessment for California

Teachers (PACT). *Issues in Teacher Education, 18*(1), 39–62.

Olafson, L. J., Schraw, G., & Vanderveldt, M. (2010). Consistency and development of teachers' epistemological and ontological world views. *Learning Environments Research, 13*, 243–266.

Pecheone, T. L., & Wei, R. R. C. (2007). *PACT technical report: Summary of validity and reliability studies for the 2003–04 pilot year*. Retrieved from http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf

Performance Assessment for California Teachers. (2007). *Appendix F: Handbook for implementing scorer training*. Retrieved from http://www.pacttpa.org/_files/Publications_and_Presentations/Appendix_F.pdf

Performance Assessment for California Teachers. (2008a). *Brief overview of PACT*. Retrieved from http://www.pacttpa.org/_main/hub.php?pageName=FAQ

Performance Assessment for California Teachers. (2008b). *Elementary mathematics rubrics 2008–2009*. http://www.pacttpa.org/_main/hub.php?pageName=Rubrics

Performance Assessment for California Teachers. (2009). *Elementary mathematics rubrics 2009–2010*. http://www.pacttpa.org/_main/hub.php?pageName=Rubrics

Performance Assessment for California Teachers. (2012a). *Elementary mathematics rubrics 2012–2013*. Retrieved from http://www.pacttpa.org/

Performance Assessment for California Teachers. (2012b). *Elementary Mathematics Teaching Event candidate handbook 2012–13*. Retrieved from http://www.pacttpa.org/

Pimm, D. (1987). *Speaking mathematically: Communication in mathematics classrooms*. New York, NY: Routledge Kegan Paul.

Porter, J. M. (2010). *Performance Assessment for California Teachers (PACT): An evaluation of inter-rater reliability* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 757340264)

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

Sato, M. (2014). What is the underlying conception of teaching of the edTPA? *Journal of Teacher Education, 65*(5), 421–434. doi:10.1177/0022487114542518

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science, 26*, 113–125.

Schwartz, R. A. (2012). *The development and psychometric modeling of an embedded assessment for a data modeling and statistical reasoning learning progression* (Unpublished doctoral dissertation). University of California, Berkeley. Retrieved from http://escholarship.org/uc/item/96j6w7xk

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

Snow, C. E. (2010). Academic language and the challenge of reading for learning. *Science, 328*(5977), 450–452.

Snow, C. E., & Uccelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 112–133). Cambridge, UK: Cambridge University Press.

Spanos, G., Rhodes, N., Dale, T. C., & Crandall, J. (1988). Linguistic features of mathematical problem solving. In R. Cocking & J. P. Mestre (Eds.), *Linguistic and cultural influences on learning mathematics* (pp. 221–240). Mahwah, NJ: Lawrence Erlbaum Associates.

Stanford Center for Assessment, Learning, and Equity. (2013). *edTPA Elementary mathemat-*

ics: Assessment handbook. Retrieved from http://blogs.oregonstate.edu/licensurefaculty/files/2014/11/elementary-mathematics-assessment-handbook.pdf

Stanford Center for Assessment, Learning, and Equity. (2015, May 1). *edTPA teaching performance assessment*. Retrieved from https://scale.stanford.edu/teaching/edtpa

Téllez, K. (2010). *Teaching English language learners: Fostering language and the democratic experience*. Boulder, CO: Paradigm.

Valdés, G., Kibler, A., & Walqui, A. (2014, March). *Changes in the expertise of ESL professionals: Knowledge and action in an era of new standards*. Alexandria, VA: TESOL International Association.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M., Hallam, P. J., Pecheone, R. L., & Moss, P. A. (2014). Evaluating the validity of portfolio assessments for licensure decisions. *Education Policy Analysis Archives, 22*(6), 1–27.

Wittgenstein, L. (1970). *Zettel* (G. E. M. Anscombe, Trans.). Berkeley, CA: University of California Press.

Wittgenstein, L. (2001). *Philosophical investigations* (3rd ed., G. E. M. Anscombe, Trans.). Malden, MA: Blackwell.

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

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modeling software*. Camberwell, Victoria: ACER Press, Australian Council for Educational Research.