PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-25

Pontificia Universidad Católica de Chile
Facultad de Educación

Assessing assessment literacy: An item response modeling approach for teacher educators

Examinando formas de hacer evaluación: Un acercamiento desde la teoría de respuesta del ítem para educadores de profesores

Brent Duckor[1], Karen Draney[2] & Mark Wilson[2]
[1]San Jose State University
[2]University of California, Berkeley

Abstract

The study articulates how to meaningfully and consistently distinguish between levels of classroom assessment knowledge among pre-service teachers using a learning progressions framework. Guided by nationally recognized principles of development and acquisition of assessment literacy, the Classroom Assessment Literacy (CAL) instrument used to identify qualitative and quantitative differences in beginning classroom assessors' proficiency estimates on a scale yields insight into new directions for modeling teacher learning progressions in this complex domain. Preliminary findings are relevant to assessment educators who are interested in diagnostic and formative uses of the CAL scales for evaluating pre-service teacher growth.

Keywords: assessment literacy; learning progressions; pre-service teacher education; item response theory; validity; reliability

Post to:
Brent Duckor, Associate Professor
Department of Secondary Education SH 436
Lurie College of Education
San Jose State University
Email: Brent.Duckor@sjsu.edu

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.5

Resumen

El presente estudio examina como distinguir de manera significativa y sistemática los niveles de evaluación de conocimiento en aula entre profesores practicantes, utilizando un marco de progresión del aprendizaje. Guiados por principios nacionalmente reconocidos sobre el desarrollo y adquisición de evaluación en aula (CAL por sus siglas en inglés), el instrumento fue usado para identificar las diferencias cualitativas y cuantitativas sobre evaluación de aula de educadores principiantes. El resultado de rendimiento en la escala presenta nuevas direcciones para modelar el entrenamiento de los profesores practicantes y sus progresiones en este complejo ámbito. Los descubrimientos preliminares resultan relevantes para educadores que utilicen evaluaciones y estén interesados en usos diagnósticos y formativos del instrumento presentado para evaluar el crecimiento de los profesores practicantes.

Keywords: progresiones de aprendizaje, teoría de evaluación, educación de profesores en formación, teoría de respuesta del ítem, validación, fiabilidad

On the eve of No Child Left Behind (2001) legislation, the nation's assessment experts called for more "assessment literacy"—for teachers, staff, and school administrators in K-12 education. For more than two decades, researchers have been interested in the topic of assessment literacy and its importance to teachers, students, and parents (Stiggins, 2001, 2002; Marzano, Pickering, & Pollock, 2001; Popham, 2004; Darling-Hammond, 2006; Gotch & French, 2014). While these experts pushed for more robust engagement with the topic, national and state policymakers in the United States sought to include clear, robust language about the role of assessment in their professional standards documents (CCSSO, 2010; NBPTS, 2012; NRC, 2010). Signaling both its importance for teachers' professional practice and more broadly for K-12 educational reform, "assessment literacy" became a clarion call.

With the emergence of the Common Core Standards movement and the large-scale assessment policy vision articulated by groups such as the Smarter Balanced Assessment Consortium in the United States, the focus on pre-service teachers' assessment knowledge and skills has come into sharper focus over the last several years. In California, for example, the professional teaching standards address several learning goals for pre-service teachers targeted on classroom assessment practices (CCTC, 2012). The revised Teaching Performance Expectations (CCTC, 2016), for example, carefully describe the set of knowledge, skills, and abilities under TPE 5 (Assessing Student Learning) that California expects teacher credential candidates to master before obtaining a license.

While the professional standards documents and state licensure bodies have been instrumental in defining what it means to be "assessment literate," other approaches and perspectives on teacher's understanding and use of classroom assessment have been neglected in the Race to the Top. One legacy of the State-led accountability and "high stakes" testing era in the U.S. has been to deemphasize the role of teacher as learner and to relegate examples of teacher expertise in classroom assessment to past educational reforms (Duckor & Perlstein, 2014).

In this article, we explore the notion of teacher learning progressions, in particular, how novice teachers make progress towards more sophisticated understanding and use of classroom assessment practices in a teacher education program (see, e.g., Shavelson, Moss, Wilson, Duckor, Baron, & Wilmot, 2010; Duckor, 2017). The concept of teacher-as-learner requires us to hypothesize continua of professional practice for pre-service teachers. Building on nationally recognized Standards for the teaching profession and the National Research Council's (2001a) report on Knowing what students know: The science and design of educational assessment, we empirically investigate how novices can approach and learn about the logic of assessment within a particular context for learning, the pre-service preparation programs in the U.S. and California in particular.

As teacher educators and educational measurement specialists, we maintain that part of acquiring literacy in the "logic of classroom assessment" involves the development of more sophisticated mental models and schema. It also involves opportunities to use mental tools, language and apply relevant assessment knowledge in different contexts and zones of professional practice (Vygotsky, 1978). Research on teacher learning progressions may help to describe and evaluate these cognitive trajectories for beginners, particularly at the intersection between field-based clinical and university-based classroom experiences. As the title of this article suggests, we are interested in the emergence and development of classroom assessment literacy for pre-service teachers who are building up their understanding of assessment as they enter the teaching profession.

The overarching questions that motivate this study are: Are pre-service teachers prepared to embrace and advance the new assessment literacy envisioned by the experts? What kinds of mental models (e.g., p-prims and misconceptions) do pre-service teachers bring to the topic of classroom assessment, for example, related to grading or scoring? How might teacher educators better understand the psychological construct "assessment literacy" so as to better prepare teacher candidates for more powerful ways of thinking about classroom assessment today? In California, new models of State-led accountability that promote continuous improvement coupled with Teacher Performance Assessments aligned with deeper classroom level assessment practices hold promise for a new, more robust notion of assessment literacy.

Background and context

Experts in educational assessment, measurement and testing agree that cognition, observation, and interpretation are essential to understanding classroom and large-scale assessment (NRC, 2001a). Each of these components comprise the logic of any assessment system: together these vertices ("topics") provide evidence to ground validation efforts, for example, to support fair and appropriate uses of data.

Building on the NRC's experts' mental model of the "Assessment Triangle" we explore the evidence for pre-service teacher learning progressions with similar components and the logic originally developed by these assessment specialists. A modified version of the NRC Assessment Triangle is depicted in Figure 1:

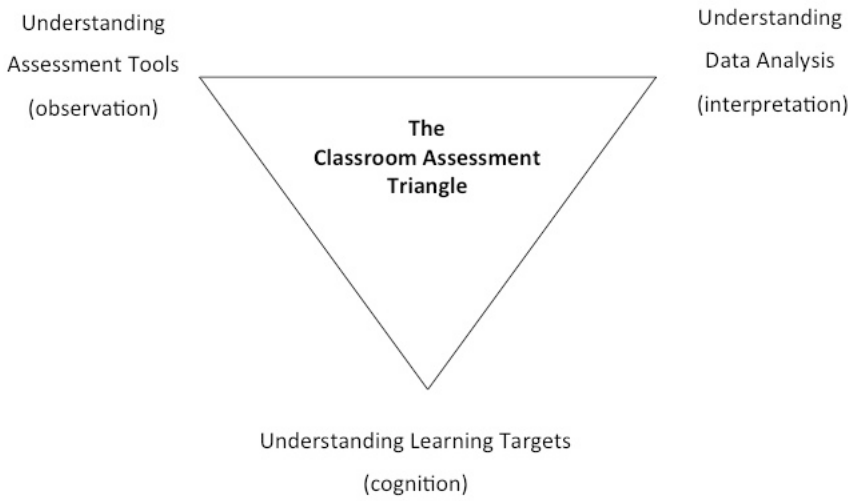

Figure 1. Classroom Assessment Triangle (adapted from NRC, 2001).

Our definition of assessment literacy draws heavily on the logic of assessment design and use depicted in Figure 1, with a particular emphasis on how teachers can and should use their own assessments in the classroom. Each vertex—"Learning Targets" "Assessment Tools" and "Data Interpretation"—represents the logic of classroom assessment around three main topics, using terms that are more accessible to pre-service teachers.

We hypothesize that knowledge of the Classroom Assessment Triangle ("CAT") has at least three distinct domains, each with a corresponding learning progression for pre-service teacher populations.

The first teacher learning progression "Understanding Learning Targets" describes a teacher's understanding of student cognition as part of classroom assessment. It defines differences in pre-service teachers' sophistication with recognizing cognitive demands, skills and levels of student understanding in classroom assessment situations. The teacher's skill with connecting learning targets to the assessment of, for example, higher and lower-ordered thinking skills is a feature of this progression.

The second teacher learning progression "Understanding Assessment Tools" describes pre-service teachers' understanding of the affordances and constraints of different assessment tools and strategies. It defines a range of understandings with item formats, modalities, and features of what experts refer to as the items design. The pre-service teacher's skill with designing and improving these assessment tools, including the anticipation of alternative scoring or sense-making procedures, is a key element of this progression.

The third teacher learning progression "Understanding Data Interpretation" describes pre-service teachers' understanding of the quality of classroom data, including making valid inferences and conclusions based on scores. The pre-service teacher's knowledge and skills with collecting and evaluating different types of validity and reliability evidence to support (or challenge) an instrument's use is a critical feature of this progression.

To operationalize these hypothesized pre-service teacher learning progressions, we examine three research questions: 1) are there distinct levels of teacher understanding of topics in a modified version of the NRC Assessment Triangle? 2) if so, how does one meaningfully and consistently distinguish between these levels of teachers' understanding on a psychometrically sound scale? and 3) what, if any, evidence for reliability and validity of scale scores is available to support diagnostic and formative uses for teacher educators?[1]

Thus, a central purpose of our research on teacher learning progressions is to measure teachers and calibrate items in three topic areas that cover the major domains of a nationally recognized framework for assessment expertise. Employing the Rasch model, we fit the data generated by a pre-post test instrument to evaluate the proposed continuum of assessment literacies among a sample of pre-service teachers. Rather than follow the expert-novice distinction common in the qualitative literature on assessment, we define the Classroom Assessment Literacy ("CAL") construct space as a potential set of teacher learning progressions which can be scaled using quantitative methods (See, e.g., Duckor, Draney, & Wilson, 2009). We also use the constructing measures approach (Wilson, 2005) to investigate the construct of "assessment literacy" itself. Our goal as pre-service educators and psychometricians is to determine to what extent, if any, meaningful variation among individual teachers exists on a clearly defined continuum of task-based performances. In the next section, we elaborate on a construct definition of "assessment literacy" that can be subject to empirical and psychometric study while embedded in pre-service program curricula at a large state university.

Analytic Framework & Methodology

The task of defining the domain of classroom assessment expertise in terms of "literacy" is challenging and likely to stir controversy.[2] Nonetheless, a review of the state and national Standards for the teaching profession reveal common features of a psychological construct that can be provisionally deemed a type of assessment literacy for teachers (CCTC, 2012, 2016; NBPTS, 2012). According to the Standards for Teacher Competence in the Educational Assessment of Students (AFT, NCME, & NEA, 1990), assessment is defined as "the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy." The 1990 Standards, of which there are seven, provide criteria for teacher competence with respect to the various components of this broad definition of assessment.[3]

While appealing to Standards documents to define classroom-based assessment literacy, our approach further aims to describe the content and structure of latent variables using a particular construct modeling approach (Wilson, 2005), one that puts a premium on evidence-centered design (Mislevy, Almond, & Lukas, 2003) and advances in item response theory and measurement.

In this Rasch IRT study, our primary goal is to construct a measure of teachers in terms of latent "proficiency" and calibrate items in terms of "difficulty" on a technically sound scale. We are also interested in the validation of scores derived from the Classroom Assessment Literacy ("CAL") instrument based on an investigation of the evidence for validity and reliability. The appropriate uses of the CAL instrument depend in large part on the stability and meaningfulness of the inferences that can be warranted based on Standards (AERA, APA, NCME, 1999, 2014).

To conceptualize our initial ideas about the structure of learning in the CAL variable, we employed the Structure of the Learning Outcome (SOLO) taxonomy. The SOLO taxonomy (Biggs & Collis, 1982) is a general theoretical framework that may be used to construct scoring or coding strategy intended to elicit a subject's (in this case, the classroom assessor's) level of cognitive sophistication with a mix of fixed choice items and written performance tasks. We use the SOLO approach to shift the focus from a hierarchy of fixed ontological stages (expert or novice) to a hierarchy of observable outcome categories (from discordant to integrative) for teachers whose expertise in a particular domain of practice is developing over time.

In the CAL framework, classroom assessors which include but are not limited to pre-service teachers are expected to draw upon at least three topics of knowledge to demonstrate proficiency with understanding classroom assessment. While we suspect that some of the proficiencies across the topics may be strongly related we nonetheless sought to carefully distinguish between each of the topics in the construct definition phase. Hence, a total of four construct maps were initially developed to represent each of the topics in the CAL framework.

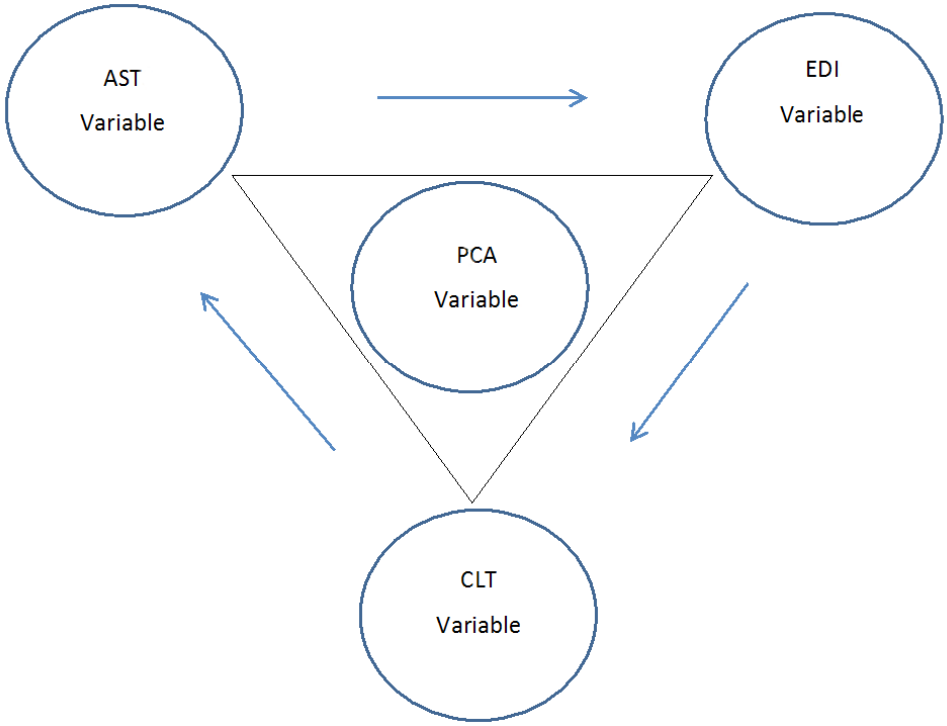

Figure 2. Relations among Domain Topics of CAL ("Classroom assessment literacy") Variables.

In the first topic, there is the Understanding Cognition and Learning Targets construct map (or "CLT variable"), which focuses on the types and quality of the representations the classroom assessor uses to define an assessment target. This variable also covers teachers' understanding of the properties, affordances, and the constraints of particular mapping procedures. The CLT variable represents the classroom assessors' skill at designing and evaluating student learning targets in terms of, for example, providing opportunities to observe for higher and lower order thinking.

The second topic is the Understanding the Assessment Strategies and Tools construct map (or "AST variable"). This variable focuses on the classroom assessor's knowledge of traditional item formats and uses, in addition to the general rules for constructing "good" items. This variable also encompasses the more sophisticated notion of items as samples from a pool that may (or may not) bear a plausible relationship to the student thinking, including but not limited to misconceptions and misunderstandings. The AST variable represents the classroom assessor's skill at designing and evaluating assessment activities (e.g., questions, tasks, items, tests, and so forth) that are aligned with the learning targets defined by, for example, the CLT variable.

The third topic is the Understanding Evidence and Data Interpretation construct map (or "EDI variable"). It includes the classroom assessor's knowledge and use of the properties of scoring strategies, which depend on purpose, context and use. This variable addresses the more sophisticated notion of rubrics, answer keys and other grading tools as methods for generating outcomes that can be used for summative, formative, and diagnostic purposes. The focus on the interpretations and uses of the student data imply a larger concern for validation, including the reliability of results. If rubrics, for example, are not aligned with content standards, taxonomical levels, or cognitive outcomes, then inferences about student learning will miss the mark. Hence, the EDI variable represents the classroom assessor's skill at designing, evaluating, and modifying scoring strategies that are aligned with elements of the CLT (student learning targets) and AST (assessment strategies and tools) variables respectively.[4]

It is important to note that each construct map is characterized by variation (or a "continuum") of performance levels with respect to both classroom assessors ("persons") and responses to tasks ("items"). The construct map articulates the expected structure of outcomes by difficulty of the items/tasks (Wilson, 2005). It provides a generalized coding scheme that will eventually be mapped back to particular scoring guides. Classroom assessors may or may not conform to these (construct) expectations, hence the map is subject to revision based on empirical findings. Figure 3 provides an example of the construct map we developed for the CLT topic domain.

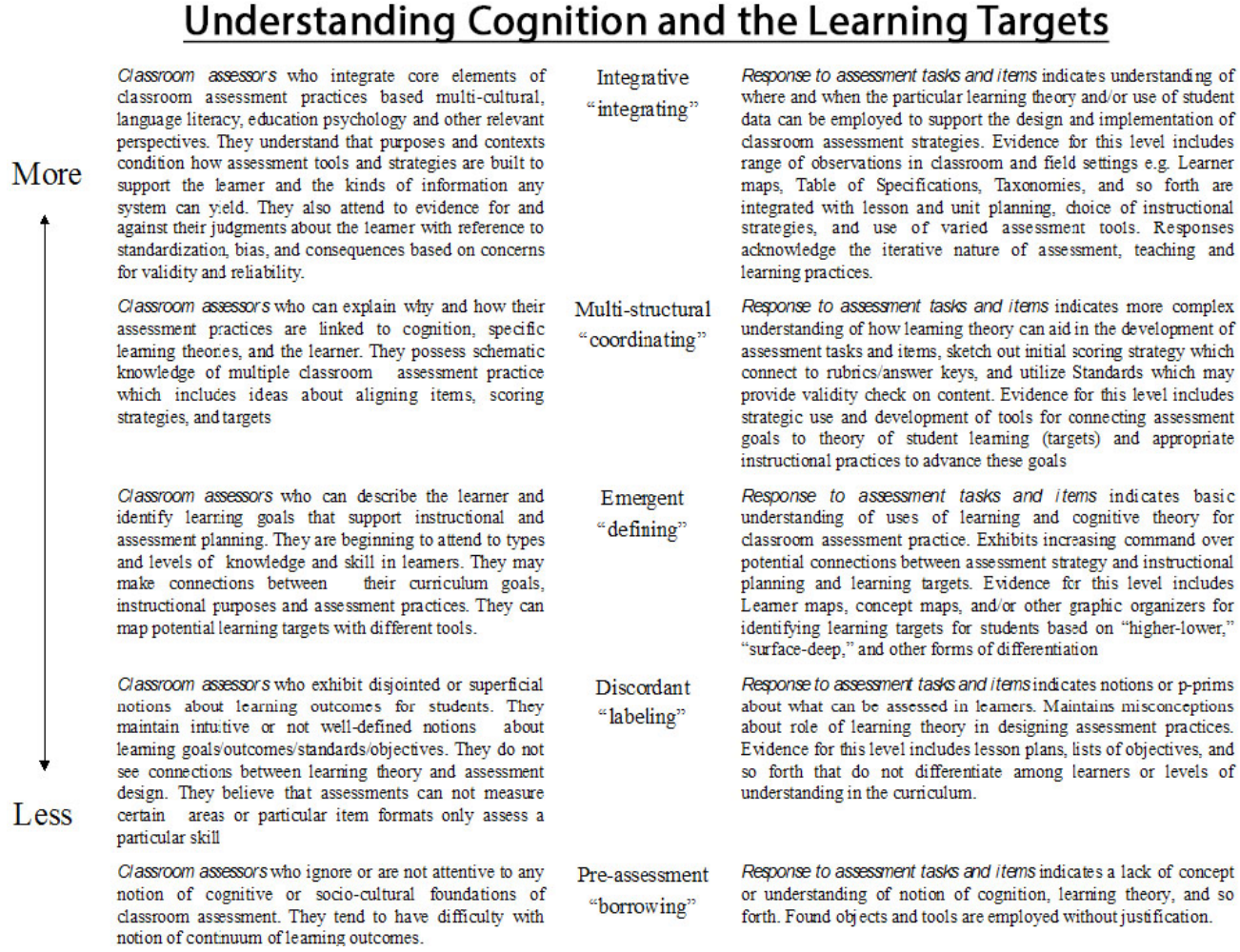

Figure 3. Construct map for Understanding Cognition and Learning Targets ("CLT" Variable").

As shown in Figure 3, at the lower extreme of the CLT construct map, we posit the existence of novice classroom assessors (i.e., those at the "pre-assessment" level) who are not yet adept at developing, analyzing and modifying different aspects of the student cognition element of the assessment triangle. These individuals tend to ignore or are not attentive to any notion of cognitive or socio-cultural foundations of classroom assessment (see, e.g., Shepard, 2000). They may possess fixed beliefs about the nature of student learning outcomes and intelligence (see, e.g., Dweck, 2010), which can contribute to difficulty conceptualizing a continuum of student performance. Their knowledge of assessment practices, principles, and "moves" is disjointed and incomplete (see e.g., e.g., Berliner, 1988; Ball & Cohen, 1999; Duckor, 2014, Duckor & Holmberg, 2017). Beginning classroom assessors at this level tend to wonder: "what does assessment have to do with teaching"? (Popham, 2007b) or might declare "Grading is how I assess" (Guskey, 2006) or believe that a "score is a score is a score" (Braun & Mislevy, 2005).

At the upper extreme of the CLT construct map, we hypothesize a group of classroom assessment experts (i.e., those at the "integrative" level) who can identify and use various mental models for representing cognitive, observational and interpretive aspects of assessment. These individuals are able to flexibly use and adapt the elements of the CAT (or mental models similar to it) while recognizing the potential affordances and constraints of a particular assessment practice. These individuals understand that purposes and contexts condition how assessment tools, procedures and strategies are designed and revised to support the learner.[5] Expert classroom assessors also attend to evidence for and against their judgments about the learner's progress with reference to concepts such as standardization ("access"), validity ("meaningfulness") and reliability ("consistency"). That is, when examining students' score data to make a judgment, they recognize the provisional nature of the inferences and the need for evidence to support evaluations of student learning.

While we are fairly confident that we have identified the extremes in our mapping of the CLT variable, we are also interested in the levels in between the classroom assessment "experts" and "novices" especially in our roles as teacher educators. We have observed that early encounters with identifying learning goals in a unit plan, for example, often leave students of classroom assessment in a "discordant" state (i.e., level 2 in Fig. 3). These persons are able to describe and label multiple goals, standards, objectives, and so forth that they wish to assess in a unit plan, for example, but they have difficulty focusing on the definition of any single, learning target for a lesson or set of lessons. Researchers (Heritage, Kim, & Vendlinski, 2008) have underscored the importance of "clarity for teachers about what comes before or after a particular learning goal." (p.4). In other words, without a mental framework such as a "learning progression" or "facets of understanding" these teachers may struggle to hold all the "moving parts" together. At this level, we observe that the teacher's cognitive mapping skills for revealing patterns in student thinking (e.g., identifying common "misconceptions" in the content area) are not very well-defined, exhaustive or ordered in particularly meaningful ways.[6]

The next level in the classroom assessment literacy progression for the CLT variable we call "emergent" (Fig. 3) in part because it refers to individuals who are emerging as competent users of a particular cognitive mapping strategy such as concept or learner maps. These classroom assessors are beginning to deal with cognitive complexity in their subject content, in part, by breaking down or chunking phenomena into multiple or "sub-targets" in their lesson and unit planning strategies. Classroom assessment practices at this level attempt to align quizzes, tests and performance tasks in the classroom with Standards or knowledge types—yielding a more substantial rationale for grading and/or awarding points. At this level of proficiency, individuals are typically able to identify the presence of potentially confounding aspects embedded in a learning target sometimes referred to as a "big idea", "central focus", "objective", and so forth.

A typical emergent response in analyzing a planning map of English Language Arts assessments, for example, begins by separating reading, writing, and spoken fluency "skills"—the beginning teacher recognizes these learning targets as potentially separate but may still struggle to define sub-targets in the writing domain such as understanding the persuasive essay and its various elements, for example, voice, structure, thesis, and/or writing conventions. Too often we observe pre-service teachers at this level reflexively attempting to apply Bloom's taxonomy and "make it fit" the problem of representing and hence assessing student learning goals and skills; or they fall back on "the students will be able to" statements about the role of a discrete assessment tool or activity in a lesson plan.

At the intermediate zone of the teacher learning progression, we have hypothesized a "multi-structural" level of proficiency in the CLT variable (Fig. 3). At this level, individual teachers are adept at explaining why and how their assessment practices are linked to cognition, socio-cultural learning theory, and the learner (see e.g., Shepard, 2000). These classroom assessors display schematic knowledge of multiple classroom assessment practices, which they coordinate when aligning items, scoring strategies, and learning targets to strengthen the chain of inference—from score to judgment. They are much less focused on grading practices and more involved in recording patterns of student responses, making connections to patterns of student understanding, and teaching towards the particular content demands of the unit topic.

Presumably teachers at this "in between" level would have sufficient knowledge to be able to pull out short-term goals for manageable chunks of instruction, coordinate formative assessment strategies, while also being able to locate the purpose of any one lesson in a trajectory of instruction that supports student learning over time (Alonzo & Gearhart, 2006). Sometimes called "master teachers" these individuals are also classroom assessment experts who have a degree of control over notions related to validity, reliability, and item bias. Many of these teachers have been involved in calibration sessions on exit exams at their schools, led school-wide discussions on standards-based grading, or may also have served as liasons to districts in the development of items, tasks and benchmark tests (See, e.g., Darling-Hammond, Ancess, & Falk, 1995).

Having qualitatively defined the structure of the proposed CAL framework, in part, by providing an example of the domain content for the CLT topic, we now turn to examine evidence for or against the overall construct theory of "proficiency" in the modified NRC (2001a) framework. Our primary interest in these developmental levels of proficiency is to better understand differences in teacher performance so as to improve the learning outcomes for both "novices" and "experts" and perhaps more importantly, for others "in between" those traditional research categories (See, e.g., Borko & Livingston, 1989; Putnam & Borko, 2000; Feiman-Nemser, 2001).

We have provisionally coded these performance levels—pre-assessment, discordant, emergent, multi-structural and integrative—to sharpen the construct theory, which represents a continuum of assessment knowledge and expertise. Utilizing nationally recognized principles of assessment design and practice, we propose that there are qualitatively important differences in both the understanding and practice of classroom assessment. We enumerate the logic of classroom assessment as recognized by the NRC rather than inventory the bevy of skills commonly outlined in professional teaching standards.

We now address the quantitative methods and data sources employed to investigate our hypotheses about the structure and functioning of these variables in a sample of pre-service teachers at a diverse, large public university that prepares single subject credential candidates in California.

Methods & Data Sources

Description of the respondents

Sample. A sample of 72 respondents was obtained from three class sections of subjects for this study. Each section consisted of pre-service teachers who participated in a post baccalaureate course offered to single subject credential candidates at a large California State University. The course is taught at the College of Education and runs concurrently with Phase II/III student teaching field experience across different middle and high school sites in Northern California. Table 1 shows the demographic characteristics of study participants by count and frequency percentage.

Table 1
Selected Sample Demographic Characteristics (n=72)

Characteristic	Count	Frequency Percentage
Female	34	47.2%
Caucasian	47	65.3%
Over 40 years old	14	19.4%
Subject credential		
Math	11	15.3%
Science	11	15.3%
Other	50	69.4%

Oral communication was the primary method of recruitment for the sample. The sample was obtained with IRB consent and data were collected as part of normal course expectations. Four exit interview items were also included with the instrument.

Instrumentation

Items design. The CAL instrument is a pre-post proficiency test designed to measure understanding and use of the NRC's (2001a) framework, with particular focus on the three topic domains related to the Assessment Triangle. The test consists of 55 items: 13 constructed response and 42 fixed choice questions. Each item is targeted on a specific domain in the CAL framework and is designed to span parts of a specific CAL construct map. An example of a constructed response item from the CLT domain is shown in Figure 4.

This item is typical of the constructed response format used on the CAL instrument. It was designed to probe the understanding of specificity, directionality and ordering of student cognitive outcomes as it relates to the task, in this case, of learner mapping. The item provides a written scenario, along with a representation of a poorly designed Learner map. There are two open-ended prompts that require a short answer. Similar to the PACT assessment task format, respondents are expected to provide a written explanation to support the use of their own classroom tools.

[1.3] A team of teachers from a local high school is developing a measure of writing proficiency. They propose to assess the following:

Proficiency with writing

Levels of student understanding	Lesson plan & instructional activities	Evidence for writing skills from assessments
Students who demonstrate mastery of writing		Final
Students who demonstrate basic proficiency by showing control over most elements of writing e.g. structure, style and voice	Genres of Writing e.g. persuasive essay, poems, science research papers, journalism articles, blogging, etc.	Midterm
Students who demonstrate limited writing skills	Uses of Voice, Grammar, Syntax, etc.	
Students who can't write	Thesis formation	Quick Write

Generally speaking, is this a good example of a Learner map? Please explain.

What advice would you give to improve it?

Figure 4. Constructed Response Item on CAL Instrument from the CLT Domain

Scoring procedure. We employed a polytomous scoring strategy for this data set. Scoring guides were used to code responses for both constructed response and fixed choice items. We developed each individual scoring guide in alignment with the CAL construct maps, which were treated as the generalized "outcome space" (Wilson, 2005). Iterating from the general to specific coding requirements, we arrived at what we refer to as scoring guides. An example is shown in Figure 5 for the CLT topic domain.

Domain Cognition and Learning Targets	Item (1.3)	Exemplar Descriptions Representing student learning outcomes with a Learner Map
Integrative	4	Understands properties of well-defined and singular learning target (construct) □ May express validity concern over multiple targets (dimensionality)
Multi-structural	3	Identifies problems with learning target (construct) definition among other things □ Offers relevant advice e.g. Use "Bloom's" to better define targets OR focus on progress on one big idea OR separate into "multiple maps" to better capture trajectory
Emergent	2	Recognizes at least one conventional, surface feature e.g. gaps, levels, types of evidence □ Offers generic advice e.g. "needs specific descriptions," "add more levels," or "more examples of student work" □ May *assume* one must apply a specific taxonomy e.g. Blooms' levels
Discordant	1	States "Looks fine" OR offers vague and misleading advice
Pre-assessment	0	No response or off topic

Figure 5. Scoring Guide Aligned with Generalized CAL Variable

These scoring guides have been designed to align with the generalized outcome space, so that the overall structure of the variable is preserved, that is, the categories used for describing levels of proficiency and item difficulty are consistent across each guide. These "exemplar" guides were primarily employed because they provide support for rater scoring protocols. All constructed response items for pre- and post-tests in this investigation were blind scored by the lead author (i.e., each response was masked to remove any personal identifiers) to reduce potential intra-rater bias.[7]

In addition to the scoring guides generated for each topic area, we also used an Ordered Multiple Choice design strategy for the fixed choice items (see, e.g., Briggs, Alonzo, Schwab, & Wilson, 2006). This allowed us to justify the award of partial credit scoring and better understand the rationale for the structure of the data. These scoring guides also allow for more flexibility in measurement model specification, for example, when exploring polytomous or dichotomous coding of data generated by the fixed choice items.

Statistical procedures

Measurement model. The choice of any measurement model is always constrained by the affordances of data (e.g., sample size, item format, and dimensionality). In this study, we employed a Rasch-family item response model to calibrate items and measure persons (Rasch, 1960; Wright, 1968; Wright and Masters, 1982). The Partial Credit Model is a polytomous version of the Rasch model (Fischer & Molenaar, 1995). It models the probability of going from level j to $j + 1$ given that the examinee has completed the step from level $j - 1$ to j. For this instrument, there are four levels and three step parameters to be estimated for each item. Formally stated in Equation 1, for the unidimensional PCM, the probability that examinee n completes step j for item i is:

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}, \quad (1)$$

where β_n is person *n*'s ability parameter, and δ_{ij} is the step parameter for the *j*th step for item *i* (Wright & Masters, 1982). This model expresses the probability of success as a function of the *difference* between the person location and the item-step location. The parameters were estimated using ConQuest software (Wu, Adams & Wilson, 1998). [8]

Statistical reports generated by ConQuest are used to describe estimates of the person and item parameters and allow for the investigation of CAL scale properties, including IRT and traditional item analyses. We also employ standard IRT analyses of item and person fit statistics to check model fit. In the next section, we report the results of validity and reliability studies conducted on the constructed response portion of the pre-post CAL instrument using these psychometric procedures.

Results

The results reported in this section are based on evidence for and against inferences about person (in this study, pre-service teacher candidate) proficiency based on CAL scale estimates. We found evidence to support our hypothesis for a unidimensional structure of pre-service teacher proficiency in understanding the NRC framework and the three domains represented by the modified Assessment Triangle (Figure 1). First, we investigated if there are distinct *levels* of teacher understanding of topics in a modified version of the Assessment Triangle. Second, we examine how to meaningfully and consistently distinguish between these levels of teacher understanding using a psychometrically sound scale, or Wright Map. Third, we present the evidence for reliability and validity of scores yielded by the CAL scale and then considered the appropriate diagnostic and formative uses of the instrument. We address the implications of these findings for the three research questions in more detail in the Discussion section.

The Testing Standards (AERA, APA, NCME, 1999, 2014) guide the evidence for reliability and validity reported in this section, and these Standards are used to establish an argument for the CAL instrument's potential interpretations and uses. Four major pieces of validity evidence (*content*, *response processes*, *internal structure* and *relations to other, external variables*) are presented here to support the meaningfulness of the scores derived from the CAL instrument.

First, our argument for *content validity* rests on (a) the development of the construct map that represents the intent to measure, (b) on the items that are designed to prompt responses, and (c) on the outcome space that is designed to value the responses according to the construct map (Wilson, 2005). The development of the CAL construct maps and items is given as an example of this content validation procedure, which occurred over a three year period. We found additional supporting evidence of "the relationship between the test's content and the construct[s] it is intended to measure" (AERA, APA, NCME, 1999, p. 11) from our literature review and several item paneling sessions with senior methods faculty, cooperating teachers, and a university Phase II supervisors.[9]

Second, we report on validity evidence based on *response processes* in order to establish evidence of "the fit between the construct and the detailed nature of performance or response engaged in by examinees" (AERA, APA, NCME, 1999, p. 12). Nearly all of the 72 respondents completed the exit interview from the CAL instrument. The overall findings from the exit interviews were positive particularly for the constructed response items: "Some of the responses for the multiple-choice items

could be made clearer such as 2.8 and 3.24." But as others wrote: "Overall, the test was very good," "covered the course material," and "The [constructed response] items were clearly worded." Based on results from the exit interviews, we conclude that the majority of respondents were neither confused nor distracted by extraneous "noise" (e.g., reading load, language complexity and so forth) that might have adversely affected their ability to respond to the items in a construct-relevant manner (Messick, 1989).[10]

Third, we report on the validity evidence for the interpretation of CAL scale scores based on the structure and functioning of the 13 constructed response items. When applying a Rasch item response model to examine the validity evidence for the *internal structure* of a scale, it is important to report the results of the weighted mean square fit and t statistics. These model fit statistics are a necessary but not sufficient guide for evaluating the scaling evidence to support intended uses, in this instance, diagnostic assessment. Our psychometric analysis of item fit statistics support the overall finding that the CAL instrument data fit the partial credit model well, which supports the validity argument for internal structure.[11]

According to the "Testing Standards" (AERA, APA, NCME, 1999), *internal structure* validity evidence refers 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" (p. 13). Based on our approach to examining the structure of a construct, a well fitting IRT model that describes qualitative differences should locate the distance between respondents and responses on a scale. We are interested in the probability of making a particular response to an item/task on the CAL scale.[12]

We used a Wright map to examine the empirical ordering of persons and items in order to compare those to our theoretical expectations based on the CAL construct maps. Figure 6 shows the distribution of respondent and item locations for the CAL scale from both pre and post-test results.

In comparing our CAL construct theory to the empirical data analysis of the CAL Wright map, we found some evidence for the banding of the item thresholds, which would be consistent with the responses to items from the same levels having similar difficulties across most items. As shown in Figure 6, the item thresholds representing the different levels of the CAL scoring guide (see Figure 5) occupy different "bands" of the scale—these are distinct except for some overlap between first and second levels. Specifically, the *prestructural* level of response to the items is represented by the first threshold (i.e., the thresholds represented by "n.1" where n ranges from 1 to 13), and spans the lower end of the scale (-6.55 to -2.66 logits). The *discordant* and *emergent* response levels are represented by the second and third thresholds (i.e., those represented by "n.2" and "n.3," respectively), which span the middle range of the scale (-2.67 to 3.25 logits). These levels represent transitional and emerging levels of proficiency in response to the CAL items. Finally, we observe that the *multi-structural* level of response, which is represented by the fourth threshold (i.e., "n.4"), covers the upper end of the scale (3.6 to 5.5 logits). Thus, we can say that as the respondents improve in CAL proficiency, they have a tendency to respond at a higher level to most of the items with more sophisticated responses.

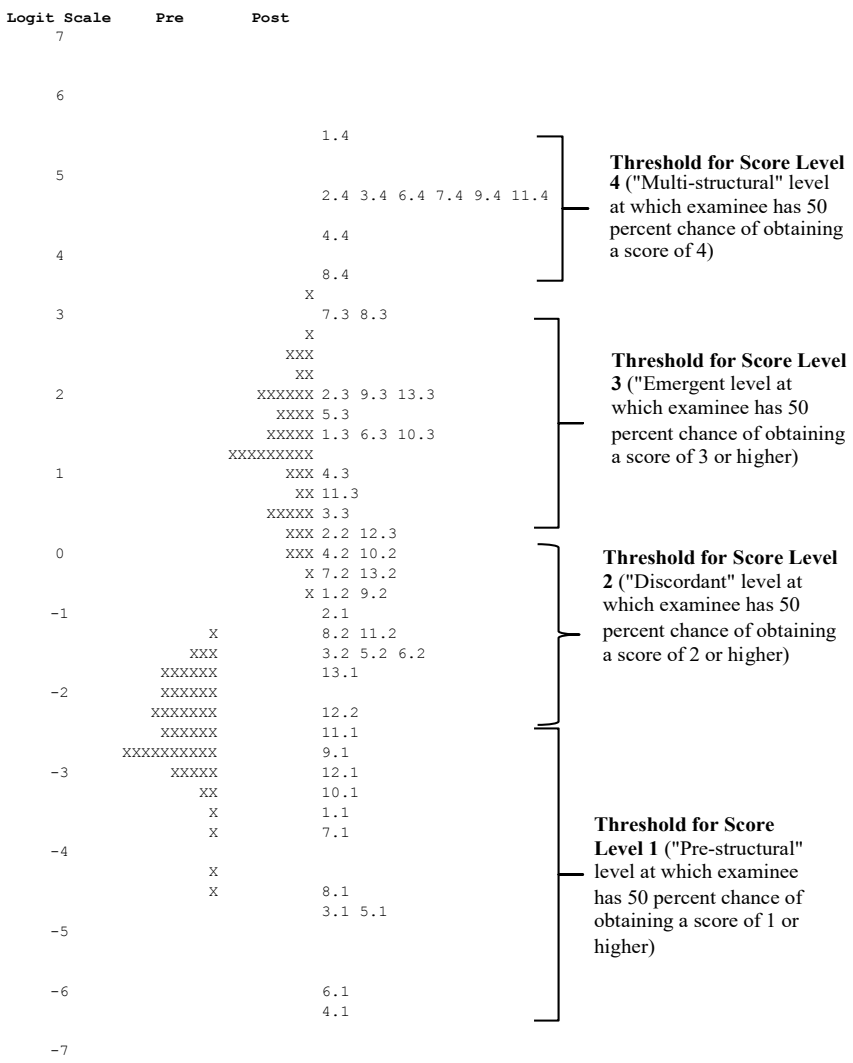

Figure 6. Wright map of Pre- and Post-Test Person Proficiencies and Item Thresholds for the CAL scale ('X' represents 1.7 cases).

The segmentation of the levels is not completely clean. Some item thresholds such as CLT 2.1 and EDI 13.1 are in an overlap region between discordant and prestructural levels of proficiency. Further investigation of these items and perhaps modification would be indicated, to see if the problem is to do with the item, or a lack of distinction between these two levels. Of course, we must exercise a degree of caution when interpreting these locations, since each is an estimate with an associated standard error.

On the whole, we observe that the levels of the CAL construct as represented in Figure 5 are mainly borne out in the empirical results in Figure 6, with some uncertainty about the boundary between the prestructural and discordant levels. In addition, the relationship between the teachers and items on the CAL Wright map indicates that the proficiencies of the respondents are covered across their entire range by the item thresholds; we did not detect either a ceiling or floor effect, which suggests the CAL instrument targets the respondents' proficiencies fairly well.[13]

Although not shown in a separate figure, the empirical results from each of the three constructs each appeared to support the theoretical expectations discussed in this study. In particular, it was found that the item thresholds for CLT, AST, and EDI sub-scales corresponded relatively well with the construct theory (i.e., hypotheses for multiple proficiencies in a uni-dimensional model). The content validity argument presented on behalf of each sub-scale provided more evidence for alignment. Although the precise structure and functioning of each of the CAL constructs requires further investigation outside the boundaries of this study, a few preliminary observations based on the results of a correlational study are presented in this section. Table 2 shows the correlations between sub-scales on the CAL instrument (for 13 constructed response items) using proficiency estimates derived from application of the partial credit model in ConQuest:

Table 2
Correlations between Sub-Scales Using WLE Estimates from the CAL Instrument

	CLT Sub-Scale	AST Sub-Scale	EDI Sub-Scale
CLT Sub-Scale		.761(*)	.817(*)
		.930	*.961*
AST Sub-Scale	.761(*)		.817(*)
	.930		*.942*
EDI Sub-Scale	.817(*)	.817(*)	
	.961	*.942*	

Note. (*) indicates correlation is significant at the 0.001 level (2-tailed test). Disattenuated correlations are italicized.

The results shown in Table 2 indicate moderately strong correlations between sub-constructs, ranging from .761 to .817. Thus, it appears that there is only slight evidence for multidimensionality of the CAL scale given the relatively strong correlations between separately calibrated Wright maps. These correlation coefficients are statistically significant, different from zero at the 0.001 level. Caution must be exercised, however, in the interpretation of correlations of IRT parameter estimates that do not account for standard errors across scales or sub-tests. In general, a correlation coefficient will be attenuated or reduced due to measurement error; thus, the estimates given in Table 2 can be thought of as lower bounds for the true correlation.

While there is room for improvement in the calibration and evaluation of the items design for this particular domain, we conclude that it provides relatively good validity evidence for the construct definition that corresponds with the general theory proposed by the CAL construct map. The results from a general item analysis add further weight to the internal structure aspect of our validity argument. Wilson (2005) notes that construct validity issues are built into the items design as well as into the construct map. One requirement is that the items be consistent with the instrument as a whole. We examine this issue for the constructed response items, in particular, by examining the mean locations of each score group on each item, which should tend to increase as the scores increase. Overall, our investigation of these different lines of internal structure evidence leads us to conclude that the items design adequately represents the construct under investigation.

The fourth source of validity evidence is based on the CAL instrument's *relations to other external variables*. Here we examined "the degree to which these relationships are consistent with the construct underlying the proposed [instrument] interpretations" (AERA, APA, NCME, 1999, p. 13). We examined the relationship between the scores on the CAL instrument and the Performance Assessment for California Teachers (specifically rubrics A6, A7, and A8 which deal with teacher candidates' understanding and use of classroom assessment). The PACT instrument is a teacher performance assessment that evaluates pre-service teachers for Tier 1 licensure. It was developed as an alternative "curriculum-embedded" assessment based on an "evidence-based system" to evaluate readiness-to-teach (Pecheone & Chung, 2007).

The results of the Pearson correlation coefficients between these PACT scores and the person proficiency estimates from the CAL instrument as a whole were positively and moderately correlated (r=.77). In fact, the correlation improved when we re-examined the relationship with the person proficiency estimates for only the constructed responses items on the CAL instrument (r=.85). This indicates that the CAL instrument scores may assess a similar set of proficiencies with understanding and using classroom assessments as the PACT instrument scores in the assessment domain.

The results of the reliability analysis we conducted are presented in terms of evidence for reasonably small standard errors of measurement. We determined this by calculating the average person standard error in comparison to the full range of person theta estimates. The ratio between them is 16.03. This means that our understanding of pre-service teachers' proficiency in classroom assessment literacy in this sample is sixteen times more precise than without the CAL instrument. We also found that the reliability coefficient values for both versions of the CAL scale were high. For the instrument as a whole, which included both fixed choice and constructed response items, the internal consistency indicators such as Cronbach's alpha (.93) were high. The reliability for the modified CAL scale, which included the 13 constructed responses items only was, in fact, slightly higher (.96).

Discussion & Limitations of the Study

This article presents the results from empirically tested hypotheses about differences in knowledge of classroom assessment among teachers, in particular, by studying 72 pre-service teachers' responses to a pre- and post-test, referred to as the CAL instrument. While we focus on the technical properties of a pilot Classroom Assessment Literacy (CAL) scale, the study has broader implications for understanding the meaning of assessment literacy.

As part of this empirical investigation, we posed three inter-related research questions about the structure and function of the construct, assessment literacy. To investigate the first question ("Are there distinct levels of teacher understanding of topics in the Assessment Triangle?") we drew from the evidence for content validity related to the CAL instrument. In contrast to previous studies and in an effort to better specify the meaning of the construct--classroom assessment literacy") we developed maps to describe the learning progression in the major topic domains. Each of these topics is aligned with the NRC (2001a) framework, which emphasizes the use and understanding of the Assessment Triangle in the science and design of K-12 assessment, a mental model drawn upon by experts.[14]

The second research question built from the first. If the CAL instrument's content appears to assess what it intends, then "how does one meaningfully and consistently distinguish between these levels of pre-service teacher understanding on a psychometrically sound scale?" Initially we approached

this investigation by fitting a partial credit Rasch item response model to the data generated by the mixed 55 item format instrument to examine our theoretical expectations about the structure of CAL proficiency. We discovered that the 13 constructed response items by themselves actually generated better model fit across all topic domains and increased the CAL instrument's reliability. That is, the constructed response items and corresponding Wright Maps, offered better estimates of the performance levels. As predicted by our theory, these empirically calibrated maps aligned well with the construct maps.

Subsequent analyses (i.e., of the association between the CLT, ATS, and EDI sub-scales) confirmed a strong degree of correlation between the three topic domains.[15] Moreover, we found a strong positive association (r=.85) between the person proficiency estimates and the PACT scores (specifically the average scores from rubrics A6, A7, and A8) for those same candidates in this study. It appears that the CAL instrument, similar to the PACT Teaching Event, is detecting teacher credential candidates' understanding and use of assessment in the classroom. Based on these statistical and psychometric IRT analyses, there is evidence to warrant meaningful, consistent distinctions between levels of proficiency—ranging from pre-assessment to more integrated understanding of the modified NRC framework—based on the CAL scale locations.

While we are confident that there is sufficient validity and reliability evidence to support particular diagnostic or formative uses of the CAL instrument, for example, in pre-service program settings, we are more cautious about further, unintended uses.[16] First and foremost, we hope to broaden the item formats and delivery platforms currently available with the CAL instrument. In particular, we see opportunities to gather more data from improved technology platforms that support a broader range of "intermediate constraint" items targeted on specific levels of proficiency. Second, inter rater reliability studies must be conducted on the constructed response items using a multi-faceted Rasch model (Linacre, 1989), particularly if usage occurs over multiple sites, raters, and occasions. Third, there are potential limits to the study of the structure and function of the CAL scales given our current measurement model specification and size of the sample (Brandt & Duckor, 2013).

Our aim in this article has been to broaden the view of the structure of assessment expertise and how it can be differentiated from more novice ways of thinking about classroom assessment. The study of differences in individual pre-service teachers' thinking about the building blocks of classroom assessment design and use is part of a longer-term research effort. It is the potential for understanding the cognitive growth of the so-called "intermediates" (neither experts nor novices) that captures our imagination as educational researchers and teacher educators. Better differentiation of potential learning progressions in the domains of instruction, curriculum and assessment for these "emergent" pre-service teachers warrants further inquiry, particularly as it relates to subject content demands. Empirical investigation of teacher learning progressions in multiple strands and domains of assessment practice—from pre-service to in-service years--is long overdue.

The original article was received on November 15th, 2016
The revised article was received on October 22 nd, 2017
The article was accepted on October 27th, 2017

References

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Melbourne, Australia: ACER.

Alonzo, A. C., & Gearhart, M. (2006). Considering learning progressions from a classroom assessment perspective. *Measurement: Interdisciplinary Research and Perspectives*, *4* (1&2), 99-108.

American Educational Research Association, American Psychological Association,
American Federation of Teachers, National Council on Measurement in Education, &
National Education Association. (1990). *The Standards for Competence in the Educational Assessment of Students.* Retrieved November 12, 2012, from http://buros.org/standards-teacher-competence-educational-assessment-students.

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.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: Author.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes and L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco, CA: Jossey Bass.

Berliner, D. C. (1988, February). The development of expertise in pedagogy. Charles W. Hunt Memorial Lecture presented at annual meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability.* Chicago: University of Chicago Press.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26, 473-498.

Brandt, S. & Duckor. (2013, June). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling, 55*(2), 148-161.

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, *86*(7), 489-497.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006) Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-64.

Brookhart, S. M. (2001). *The Standards and classroom assessment research.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. (ERIC Document Reproduction Service No. ED451189)

Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Corcoran, T.B., Mosher, F.A., & Rogat, A.D. (2009). *Learning progressions in science: An evidence-based approach to reform*. (CPRE Report). Philadelphia, PA: Consortium for Policy Research in Education.

Council of Chief State School Officers. (2010, July). Interstate Teacher Assessment and Support Consortium (InTASC) *Model Core Teaching Standards: A Resource for State Dialogue (Draft for Public Comment).* Washington, DC: Author.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education*, *57*(2), 120-138.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work.* New York: Teachers College Press.

Duckor, B. (2005, May). *Thinking about the act of measuring: The development of a theory of the construct*. Individual poster presented at the 2nd annual meeting of the Center for As- sessment and Evaluation of Student Learning Conference, Santa Rosa, California. Available from Center for Assessment and Evaluation of Student Learning at http://www.caesl.org/ conference2005/ brent_sm.pdf

Duckor, B. M. (2006). *Measuring measuring: An item response theory approach*. (Doctoral dissertation, University of California, Berkeley, 2006). 345 pp. Advisor: Wilson, Mark R. *UMI Dissertation Abstracts (ProQuest).*

Duckor, B., Draney, K. & Wilson, M. (2009). Measuring measuring: Toward a theory of proficiency with the Constructing Measures framework. *Journal of Applied Measurement*, *10*(3), 296-319.

Duckor, B. (2014, March). Formative assessment in seven good moves. *Educational Leadership,* 71(6), 28-32.

Duckor, B., & Perlstein, D. (2014). Assessing habits of mind: Teaching to the test at Central Park East Secondary School. *Teachers College Record*, 116(2), 1-33.

Duckor, B. (October, 2017). *Linking formative assessment moves with high leverage instructional practices: Rethinking translation, application and practice of classroom assessment with a learning progressions framework.* Invited speaker at Graduate School of Education, Peking University, Beijing, China.

Duckor, B., & Holmberg, C. (2017). *Mastering formative assessment moves: 7 high-leverage practices to advance student learning*. Alexandria, VA: ASCD.

Dweck, C. S. (2010). Mind-sets and equitable education. *Principal Leadership*, *10*(5), 26–29.

Feiman-Nemser, S. (2001a). From preparation to practice: Designing a continuum to strengthen and sustain practice. *Teachers College Record 103*(6), 1013-1055.

Feiman-Nemser, M. (2001b). Helping novices learn to teach: Lessons from an exemplary support teacher. *Journal of Teacher Education*, *52*, 17-30.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Gotch, C.M., & French, B. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2) 14–18.

Guskey, T. R. (2006). Making high school grades meaningful. *Phi Delta Kappan*, *87*(9), 670-675.

Heritage, H.M., Kim, J., & Vendlinski, T. (2008). Measuring teachers' mathematical knowledge for teaching (CSE Technical Report). Los Angeles, CA: Center for the Study of Evaluation and National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., & Baker, E.L. (2005). Making benchmark testing work. *Educational Leadership, 63*(3), 48-54.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Lortie, D. (1975). *Schoolteacher: a sociological study*. Chicago, IL: The University of Chicago Press.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment validity and reliability. *Mid-Western Educational Researcher, 13(4),* 29-35.
Mertler, C. A. (2003). *Classroom assessment: A practical guide for educators.* Los Angeles, CA: Pyrczak.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to Evidence Centered Design*. CRESST Technical Paper Series. Los Angeles, CA: CRESST.
Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of inter-individual and intra- individual variability: A theoretical and empirical vindication of devel- opmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development.* Dordrecht, the Netherlands: Kluwer.
National Board for Professional Teaching Standards. (2012) *The five core propositions*. Retrieved from the NBPTS website: http://www.nbpts.org/the_standards/the_five_core_propositio
National Research Council. (2001a). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J.W. Pellegrino, N. Chudowsky & R. Glaser, (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.
National Research Council. (2001b). *Tests and teaching quality*. Washington, D.C., National Academy Press.
National Research Council. (2010). *Preparing teachers: Building evidence for sound policy.* Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
Nitko, Anthony, J., and Brookhart, Susan M. (2006) *Educational assessment of students*, 5th Edition. Upper Saddle River, New Jersey: Merrill.
PACT. (2007) *A brief overview of the PACT assessment system*. Retrieved from the PACT website: http://www.pacttpa.org/_main/hub.php?pageName=Home
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 on March 17, 2012 from http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf
Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6(1),* 21-27.
Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: a national survey. Educational Measurement: Issues and Practice, 12(4), 10–12.
Popham, W. J. (1997). What's Wrong—and What's Right—with Rubrics. *Educational Leadership, 55*(4), 72-75.
Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests*. Boston: Allyn and Bacon.
Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82-83.
Popham, W. J. (2007a). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89(2),* 146-150.
Popham, W. J. (2007b). *Classroom Assessment: What teachers need to know* (5th ed.). Boston: Allyn & Bacon.
Popham, W. J. (2008). What's valid? What's reliable? *Educational Leadership, 65*(5), 78-79.

Putnam, R.T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* [Reprinted by University of Chicago Press, 1980].

Shavelson, R.J, Moss, P., Wilson, M., Duckor, B., Baron, W., & Wilmot, D. (May, 2010). *The promise of teacher learning progressions: Challenges and opportunities for articulating growth in the profession.* Individual paper presented at the Teacher Learning Progressions symposium for Division D-Measurement and Research Methodology, American Education Research Association, Denver, Colorado.

Shepard, L.A. (2000). Role of learning in an assessment culture. *Educational Researcher*, *29*(7), 4-14.

Stiggins, R.J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement and Practice*, *20*, 5-15.

Stiggins, R.J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, *83*, 758-765.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238-245.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. (Edited by M. Cole, J. Scribner, V. John-Steiner, & E. Souberman). Cambridge, MA: Harvard University.

Wiggins, G. (1998). *Educative assessment*. Designing assessments to inform and improve student performance. San Francisco, CA: Jossey Bass.

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

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716-730.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325-332. Tokyo: Springer-Verlag.

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

Wu, M., & R. J. Adams. (2012). Properties of Rasch residual fit statistics. Journal of Applied Measurement, 14, 339-355.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). ACER*ConQuest* [computer program]. Hawthorn, Australia: ACER.

Endnotes

1 By "distinct" levels we mean empirically distinct based on calibrated score data. This leaves open the question of whether these levels represent any stronger construct theory e.g., developmentally distinct levels between persons. Relatedly, this study makes the assumption that individual teachers are likely to progress through successively more sophisticated levels of within-person differences in sophistication of understanding; the data is cross-sectional and cannot support claims regarding these differences over time. See, e.g., Molenaar, Huizenga, & Nesselroade (2003).

2 Stiggins (1995) states that "Assessment literates know the difference between sound and unsound assessment. They are not intimidated by the sometimes mysterious and always daunting technical world of assessment" (p. 240).

3 A few research studies have been conducted over the past 20 years that have addressed one or more of the seven Standards for Teacher Competence in the Educational Assessment of Students (AFT, NCME, & NEA, 1990). These studies often involve survey instruments that cover a broad range of practices—from developing valid grading procedures to recognizing unethical or illegal practices (Brookhart, 2001; Plake, 1993; Plake, Impara, & Fager, 1993; Campbell, Murphy, & Holt, 2002). Campbell et al. (2002) administered the Assessment Literacy Inventory (ALI) to 220 undergraduate students following course in tests and measurement. The course included topics such as creating and critiquing various methods of assessment, discussing ethical considerations related to assessment, interpreting and communicating both classroom and standardized assessment results, and discussing and evaluating psychometric qualities (i.e., validity and reliability) of assessments. Mertler (2000, 2003) used the Classroom Assessment Literacy Inventory (CALI) to compare pre- and in-service teachers "assessment literacy." Standard and total scores for the two groups of teachers were compared by conducting independent-samples t-tests (a = .05). Examination of the results revealed that significant differences existed between the two groups for scores on 5 of the 7 Standards, as well as for the total scores. Pre-service teachers performed highest on Standard 1—Choosing Appropriate Assessment.

4 We have also identified a potential fourth domain, Understanding the Principles of Classroom Assessment (PCA), which addresses the interconnectedness of each of the three topic domains, particularly with respect to notions of validity and reliability as quality control checks on both classroom-level, district and state-wide assessments. The map represents a global proficiency with understanding the Assessment Triangle as a method of inquiry to evaluate the claims of other assessors about student progress. The map also considers how teachers communicate to other stakeholders including parents, colleagues, administration and the students themselves about the science and design of assessment more broadly. The PCA domain is not explored in this study. For a more general consideration of PCA, see, Duckor & Holmberg (2017).

5 We leave open in the study the question of whether CLTs at the higher levels must be domain-specific. The skills and knowledge for assessing history or algebra likely require more differentiated, discipline-specific more complex "levels" and perhaps new construct maps (See, e.g., Wilson, 2009).

6 A typical discordant response to analyzing an assessment task in mathematics, for example, will conflate reading literacy with procedural fluency—the teacher wants to evaluate both learning targets but has not yet figured out how to represent either one in its fuller complexity.

7 Since the lead author served as the single rater in this study, there is potential for confounding intra-rater leniency/severity effects with the item/step difficulty and fit statistics. Further inter-rater reliability studies should be conducted for score data uses that involve consequential or programmatic decisions. There is sufficient reliability evidence for the current intended uses of the CAL instrument i.e., for classroom level diagnostic and formative assessment purposes.

8 The particular method for person estimation used in all analyses in this article is the Weighted Likelihood Estimate (WLE). This method provides the best individual estimates with the least bias (Wu, Adams & Wilson, 1998).

9 Building from previous research on the structure of educational measurement proficiency (Duckor, 2006; Duckor, Draney, & Wilson, 2009), interviews with content experts, university methods faculty, and course instructors in the credential program were conducted. This grounded theory approach yielded a picture of possible learning progressions, which was later turned into an initial set of CAL construct maps. These maps were revised after several instrument pilot-testing phases, in which some items were excluded in favor of others to better capture levels of performance, especially in the mid-range. We then triangulated the item responses derived from the pilot CAL instruments with an examination of student work from the EDSC 182 course, which led to further improvements in the CAL construct maps, in particular the CLT map focused on understanding student cognition and representing learning targets.

10 Most of the confusion from respondents was directed at the fixed choice items; their comments focused primarily on the distractors that "did not have one clear answer" or provided "confusing options." One respondent wrote: "I would make some of the multiple-choice answers more finite. I always feel like I am answering incorrect when there are so many plausible answers…[these questions] feel like tricks." The respondents gave detailed feedback on the instrument, including suggestions for improvement. Most of these suggestions focused on reducing the time required to complete the instrument. One student teacher commented on the testing principle embodied in the instrument: "The test is too long, past the point of adding to the test's reliability."

11 The results of the item fit analysis support the overall finding that, at the item level, the CAL instrument data fit the partial credit model reasonably well. We examined weighted mean square fit statistics (Wright & Masters, 1982) for the item step parameters, which indicate good overall fit using the interpretive framework ($.75 < MNSQ < 1.33$) developed by Adams & Khoo (1996). Only one item, ATS 2.2 (score category 3) appeared to show model misfit from a statistical standpoint. The weighted mean square value for this item (1.28) is greater than one, indicating that the observed variance is greater than expected, which given the t statistic (2.4) may not be due to chance alone. It should be noted that, given given the small sample size used in the study compared with the large number of estimated parameters, it is possible that some parameter estimates, and hence fit statistics may be inflated by Type I error rates. For discussion, see, e.g., Wu & Adams (2012).

12 A Wright Map is a useful visual tool for depicting these relationships ("distances") between person proficiencies and item difficulties on a single continuum (the "logit scale"). In the Wright Map in Figure 6 we interpret the distances between teachers "X"s and thresholds "i.k," where i indicates the item and k indicates the threshold (See, e.g., Wilson & Draney, 2002). First, if a particular teacher is at the same location as a CAL test item threshold, say threshold 2, then would mean that that the teacher is likely, for example, to score a "2"or below on that item with 50% probability. Second, a teacher with more proficiency (higher X would have a greater than 50% probability of scoring above a "2" on that particular item. Third, a teacher with less proficiency (lower X would have a greater than 50% probability of scoring below a "2" on that particular item.

13 We note that our current items design does not yet provide sufficient opportunities for respondents to demonstrate an integrated level of understanding on the CAL instrument, which we address in the discussion section.

14 Based on literature review and informal interviews conducted with student teachers, cooperating teachers, and several university Phase II supervisors over a two year period, we established that these NRC-aligned constructs could serve as the basis for embedding a test instrument in the pre-service curriculum while meeting larger program and single subject certification goals. Thus, the content validity argument advanced in this study rests upon the degree of relationship between these construct maps, the items designed to prompt teacher responses, and the scoring strategies for categorizing outcomes derived from the instrument, each of which are embedded in a particular curriculum.

15 The strong (disattenuated) correlations between the three variables may offer evidence against the argument for treating these teacher learning progressions as distinct sub-dimensions for the purposes of testing and evaluation. Nonetheless, from an instructional and pedagogical standpoint, teacher educators and their student-teachers may benefit from the analytical distinction, particularly in classroom assessment courses focusing on unit test, quiz, and other performance task designs. See, e.g., Nitko & Brookhart (2006).

16 Research on measuring teachers' "assessment literacy" should, where possible, indicate the reliability of the instruments (e.g., surveys, observation tools, tests, and so forth) used to make claims about, in this case, assessment literacy among populations of pre-service teachers. In our preliminary review, we found that the CAL scale and sub-scales have reasonably good reliability properties. Moreover, the four types of validity evidence gathered in the study each point towards a conclusion: the relationships among test items and test components conform to the construct on which the proposed CAL instrument score interpretations are based.
For a comprehensive review of different "assessment literacy" instruments and the technical evidence to support use. See, Gotch & French (2014).