The International Journal of Educational and Psychological Assessment
January 2012, Vol. 9(2)

Development and Validation of Indicators of Teacher Proficiency in Diagnostic Classroom Assessment: A Mixed Methods Study

Madhabi Chatterji
Teachers College, Columbia University
U.S.A.

Abstract

The purpose of this study was to derive and validate a domain of teacher proficiency indicators in formative, diagnostic classroom assessment, and to demonstrate the utility of the overall domain and sub-domains as a framework for designing future instruments. The domain was envisaged as a behavioral-attitudinal construct and derived using an emergent, two-stage design with mixed methods. The resulting domain and sub-domains suggest a theory on a continuum of teacher change in classroom assessment practices. Implications for future research on teacher development in diagnostic assessment skills in the U.S. and international education settings, are discussed.

Keywords: Teacher assessment, professional development, diagnostic assessment, formative assessment, classroom assessment monitoring.

Introduction

Currently, there is a dearth of models and tools to support teacher development in diagnostic, formative classroom assessment in both the U.S. and international education settings. In particular, there is limited research focusing on how teachers could incorporate updated knowledge from the cognitive sciences and learning theory into their existing repertoires of classroom instruction and assessment practice (Pellegrino, Chudowsky, & Glaser, 2001). Studies suggest that actual practices fall far behind the rhetoric of "formative assessment", with the typical teacher still viewing assessment primarily as a means for assigning student grades and lacking the knowledge necessary for competent diagnosis of learner needs, accompanied with formative strategies (Shepard, 2006).

Internationally, the research continues to show that schools are "beset with problems and shortcomings" when it comes to implementation of formative assessment in classrooms (Black, Harrison, Lee, Marshall, & Wiliam, 2003, p. 10). While teachers may engage in some degree of pupil assessment, individualization, and re-teaching, they still lack adequate skills in ongoing, diagnostic analysis of learner needs and mediation techniques (Erickson, 2007; Macintyre Latta, Buck, & Beckenhauer, 2007). To compound the problem, institutional pressures for accountability-oriented testing, coverage of curriculum, and a summative assessment culture in schools serve as threats to "formativity" in classrooms (Chatterji, Koh, Choi, & Iyengar, 2009; Erickson, 2007, p. 190).

Research on how teachers learn best and conditions that support changes in teacher practices suggests that there should be a well-conceived model of "ideal professional practice" as a guide (Shepard, 2006, p. 641). At the same time, teachers must be allowed sufficient freedom to exercise their professional judgment to modify the recommended strategies and models in ways that best facilitate their classroom goals, permitting meta-cognition, reflection, and assimilation along the way (Lave & Wenger, 1991; Putnam & Borko, 2000).

Currently, there are no formal tools to foster teacher development in diagnostic and formative classroom assessment models that combine the latest information from cognitive psychology and assessment, with best practices in pedagogy and instructional design. The present endeavor was undertaken to address this void.

The purpose of this study was to derive and validate a domain of teacher proficiency indicators in formative, diagnostic classroom assessment, and to present the overall domain and sub-domains as a framework for designing future instruments. The work began with a deductive analysis of a selected body of literature. In the second stage, findings were broadened, and cross-validated with a case study.

The case data were gathered from a selected teacher and her classroom of 16 elementary level students at a New York school. The teacher was a participant in a 2-year research and development project involving a specific approach to formative, diagnostic assessment called *Proximal Assessment for Learner Diagnosis* (PALD) (Chatterji et al., 2009). As a project participant, the teacher received formal training in diagnostic strategies rooted in the theory underlying the PALD model. For two consecutive years, she employed diagnostic assessment practices voluntarily with her class during mathematics instruction. At the end of the second year, her class was found to demonstrate gains on both internal and external achievement measures. Simultaneously, the frequency of diagnostic practices engaged by the teacher, was also observed to increase. These findings provided a compelling rationale for a deeper examination of the case.

Research Questions

The aim of the first stage of the work was to derive, compile, and organize preliminary indicators of a *PALD Proficiency* domain, using a broad but relevant literature review. Accordingly*,* the initial question was theoretically-oriented:

(1) Based on the literature, what are some observable (verbal and non-verbal) behaviors, attitudes, or practices that an expert teacher would likely display when using diagnostic assessment strategies effectively with students?

The second set of questions were formulated to compare the theoretical indicators against indicators drawn from the selected teacher's actual assessment applications with her class. The questions guiding the second stage were the following.

(2) On a day-to-day basis and in real time, how does a trained teacher engage in formative assessment and diagnostic practices in her classroom? What does the teacher say or do in her classroom that is consistent with the theoretical literature on which the PALD model is founded? In what ways, and under what conditions, do the teacher's assessment practices depart from that theoretical model?

(3) Is there a "theory of PALD practice" that is grounded in the data that suggests ways in which the teacher assimilated diagnostic assessment

approaches within her daily teaching repertoire? If so, what is the observed continuum of teacher change in diagnostic assessment practices?

Theoretical Framework for Domain Specification and Assessment Design

This section provides the theoretical background to support the methodology adopted to answer the questions. To achieve complementarity using a mixed methods approach (Greene & Caracelli, 1997), traditions in educational and psychological measurement were combined with selected methods in qualitative inquiry.

Structure of Behavioral and Attitudinal Constructs

Since *PALD Proficiency* was conceptualized as behavioral-attitudinal construct, a first challenge dealt with identification of a theoretical taxonomy to help organize, classify, and label indicators in the domain. The psychological literature on the structure of attitudinal constructs dating back to the 1960s (Rosenberg & Hovland, 1960) provided support for conceptualizing the *PALD Proficiency* construct in terms of two principal dimensions, reflecting respectively, teacher *Beliefs* in diagnostic assessment approaches, and teacher *Practices* or behaviors consistent with such approaches. Rosenberg and Hovland's tripartite taxonomy suggested that a person's underlying attitude towards an object, such as the PALD approach, could be manifested in one of three ways: (1) *Beliefs/Opinions*; (2) *Behaviors* (3) or *Feelings*. *Beliefs* represent what individuals perceive they know or hold to be true about some attitudinal object. *Behaviors* represent what individuals might actually do that reflects their underlying stance about that object. *Feelings* represent emotions evoked in individuals in relation to the object.

More recent attitudinal literature indicates that a *Belief* could also represent one's *self-efficacy* in a given proficiency domain (Bandura, 1997). *Self-efficacy* refers to a belief in one's own capacity to succeed in particular types of tasks, such as, diagnostic assessment in classroom teaching contexts. Research evidence shows that as one becomes better at performing tasks in an area, one's belief in one's own ability to succeed with similar tasks in the future, also increases. Individual beliefs in terms of self-efficacy in a domain are positively correlated with actions and behaviors (Bandura, 1997).

The *Belief* and *Behavior (Practice*) dimensions were selected for sorting and coding teacher indicators within the *PALD Proficiency* domain as they were a natural match for the construct. Given the earlier documentation of assessment challenges faced by teachers in already demanding school environments, teacher *self-efficacy* in implementation of diagnostic assessment was added as a relevant component of the broader domain of *PALD Proficiency*. The *Feeling* dimension of attitude was considered less pertinent to a proficiency construct in professional development contexts for teachers.

Traditions in Domain Specification and Assessment Design

Once a taxonomy was identified, the *PALD Proficiency* domain was specified using domain-sampling techniques following long-standing traditions in educational and psychological measurement (AERA, APA, & NCME, 1999; Chatterji, 2003; Crocker & Algina, 2006; Nunnally & Bernstien, 1994). Behavioral-attitudinal constructs are not always directly observable and must be inferred through indirect, but otherwise observable indicators that need to be identified when defining a domain.

By tradition, there are three main steps in applying the domain-sampling technique. They involve: (1) identifying demonstrable behaviors or observable indicators of the trait of interest from existing literature or documentary sources – such as, words, responses, or actions of teachers using diagnostic assessment strategies; (2) selecting or writing items matched to the domain indicators using established item-writing guidelines; and finally, (3) sampling a sub-set of the prepared items or tasks to assemble an assessment instrument that may be scored, scaled, and applied in research or practice contexts.

When applied to the design of standardized achievement tests, two steps are typically added to the above process. There is a "content-validation" step, involving external, expert reviews of the domain and items to check for content representativeness and relevance. This is followed by empirical field-tests and psychometric evaluations of items and the instrument as a whole so as to ensure construct validity, reliability, and overall quality of measures for the intended applications (Kane, 2006; Schmeiser & Welch, 2006). The domain specification and assessment design processes are iterative, so that the instruments and domain may be progressively refined to attain optimal levels of validity and utility with targeted populations, in the intended contexts of use (Chatterji, 2003).

Case Study and Grounded Theory Methodology

A departure from the domain-sampling tradition in this study involved the use of a case study and qualitative data sources for specifying a portion of the *PALD Proficiency* domain during the second iteration. Case study methods involve in-depth examinations of some bounded unit, as in the single classroom that is the focus of the present study (Mabry, 2009; Stake, 1997). Data from case studies are primarily descriptive and obtained through direct, inductive methods. Both qualitative and quantitative data may be generated, gathered via multiple sources including participant interviews, observations, field notes, documents, test scores, and other artifacts (Mabry, 2009).

Typically, qualitative data collected as a part of case studies are coded and analyzed to derive prevalent themes and patterns. When data are collected as narrative or free-flowing text, grounded theory analysis procedures provide a systematic series of steps for coding segments of the information to generate thematic categories (Strauss & Corbin, 1998). The thematic categories can then be positioned within a theoretical structure and inter-connected with each other to reveal a pattern or tell a story (Creswell, 2003). A "theory" may thereby emerge from the coded data.

Case study methods accompanied by grounded theory techniques were useful methods in the present effort for two reasons. Several forms of qualitative data helped obtain deeper understandings of classroom and teacher interactions related to diagnostic assessment "in action" under real-time conditions. The data also bolstered the theoretically-derived indicators, making possible a more authentic and comprehensive representation of the domain.

Methods

Stage One: Domain Specification based on a Literature Review

Literature sources were compiled from diverse disciplinary bases and were sorted first by primary area, namely, educational and cognitive psychology, developmental psychology, educational assessment, and pedagogy. Each article was then subjected to several readings to extract themes (main ideas) and sub-themes. The themes were classified using the selected taxonomy as falling under either the *Belief* or *Practice* categories or as applicable to both. Logical inter-connections among these categories were mapped.

Each main theme was treated as a general indicator. Sub-themes were subsumed under these as specific indicators. Indicators were labeled, clustered logically into groups, and organized hierarchically within the domain (see Chatterji, 2003). This procedure yielded four preliminary *Belief* sub-domains falling under the larger *PALD Proficiency* domain, each defined by several specific indicators.

Stage Two: Domain Indicators from the Case Study

Data sources. The study next extracted indicators of *PALD Proficiency* directly from the teacher and her class. Data were gathered from the following three sources:

(1) The teacher's documented work while teaching a mathematics unit that was designated as a PALD unit in the larger project (long division). The data included lesson planning notes, assessment artifacts, and records of reflective discussions with the project's assessment coach collected during a 2-day PALD lesson-demonstration and coaching session in Year 1.

(2) A 30-minute classroom observation of the teacher's assessment activities while teaching a Non-PALD unit independently in Year 2. Verbatim, running records were taken of events as they unfolded in the classroom and coded using grounded theory methods (See the Appendix for an illustration).

(3) Journal records kept by the teacher of assessment applications in another mathematics unit (decimals and fractions). This record was a part of the teacher's end-of project portfolio. It was identified for

analysis as it had a full complement of data on student performance corresponding with documentation of the teacher's assessment activities.

Case selection. The case was selected based on four criteria: the teacher participant had attended all PALD workshops and coaching sessions; students in the class had been exposed to PALD-trained teachers for two continuing years (in both grades 5 and 6); there was little or no attrition of students with the class demonstrating mathematics gains from grade 5 to the end of grade 6; and teacher data had been formally gathered on multiple occasions in grade 6 using at least *two* different data sources (classroom observations, observations made during coaching-demonstration sessions, or teacher journal records). The teacher in the case (*LA*) and her class of 17 students met all these criteria. One student had been lost due to mobility since grade 5, yielding 16 students.

Teacher and class characteristics. *LA* is female and White. At the time of the study, she was certified to teach at the elementary level, held a Master's degree in education, and had 9 years of teaching experience. The teacher taught in a school with a strong achievement history in the district, but with a high minority enrollment and designation as a Title 1 district.

Her student group of 16 consisted of 3 (19%) Asian, 9 (56%) Black, 3 (19%) Hispanic, 1 (6%) White children, whose birth dates fell between January, 1995 through March, 1996. Of these, 7 (44%) were girls. None were coded as English Language Learners or Disabled on the school district database, but 9 (56%) were on the free/reduced lunch program, an indicator of poverty. None were classified as gifted or exceptional.

Observed student achievement gains. At the end of grade 5 (Year 1), the class' mean raw score on the PALD unit test (long division) was 23.72 (Standard Deviation (*SD*) of 10.67). Using the pooled sample of all fifth graders in all PALD and non-PALD classes from four schools that participated in the project, the grade 5 mean placed the class .18 *SD* units *below* the mean for the pooled sample.

Following grade 6 (Year 2), the class' mean score was 32.12 (*SD*=14.66), with the class now placed +.59 *SD* units above the pooled sample of 6th graders in Year 2. In terms of gains in *SD* units, the class' mean gain registered at +.78 (*d* =+.59 – (-.18)), a statistically significant change based on a correlated means *t*-test (*p*=.011). This pattern of gain was observed in another PALD unit (geometry), as well as on the New York state's standardized mathematics test at the end of grade 6.

Changes in teacher's assessment practices. Correspondingly, *LA* also showed higher frequencies of diagnostic assessment practices with her class, consistent with the PALD model. Specifically, in 30-minute classroom observation sessions, the frequency of PALD behaviors during teaching increased from 19 at the end of the first year, to 21 in mid-year (Year 2), to 38 at the end of that same year (Year 2). Increases were documented in higher use of paper and pencil assessments for diagnosis during instruction, probing of students to check for deeper understanding, providing immediate feedback to individuals and small groups when gaps were identified, and responding to content-related questions from students.

Analytic procedures. Once the case was selected and data gathered, text data were coded sentence-by-sentence to extract key themes and sub-themes via a grounded theory analysis (Strauss & Corbin, 1998; Creswell, 2003). The Appendix demonstrates the coding methodology with data excerpts of a classroom running record. Overlapping themes were collapsed or re-worded to reduce redundancy. Unique themes that fell outside the theoretical PALD model were identified and re-organized where necessary. (T)eacher and (S)tudent action or interaction vignettes were identified representing particular practice indicators.

Cross-Validation and Triangulation

Once results were tabulated from both stages, the general and more specific indicators of PALD *Belief* and *Practice* were compared against each other and verified for consistency. Cross-validated indicators were retained. Based on frequency of occurrence or consistency with the literature, unique indicators were either added to the domain or rejected as idiosyncratic.

Preliminary Item Pool

As a final step, samples of items were written matched to domain and sub-domain indicators, using standard guidelines for the design of closed-ended, self-report or behavior rating scale questions (Fink, 1995). In the items shown, *Beliefs* may be self-rated using a 5–point Likert scale: *Strongly Agree, Agree, Uncertain, Disagree, Strongly Disagree* response options. Possible answer options for self-rated items in the *Practices* domain could be: *About once a week, About once in two weeks, About once a month*, and *Less than once a month.* Answer options for a behavior-based tool rated by an observer could be: *Observed very consistently and regularly* (once a week or more), *Observed some of the time with consistency* (once in two weeks or so), *Observed occasionally and erratically* (once a month or so), *Observed very rarely or Not observed at all* (less than once a month).

Results

Preliminary PALD Proficiency Domain: Belief Indicators

Tables 1-2 present the initial *PALD Proficiency* indicators extracted from the literature review. Table 1 summarizes the general and specific indicators drawn from the literature sources discussed next. The first set of indicators are formulated to tap into the teacher *Belief* dimension under the overall construct of *PALD Proficiency* in fostering learner development. There are four sub-domains (1.1-1.4, Tables 1-2): *Belief* in cognitive modifiability and continuous human development in a domain; *Belief* that proximal, timely assessment and mediation bridges learning gaps; *Belief* in student-centered, outcome-driven pedagogy; and *Belief* in self-capacity to conduct diagnostic and formative assessment in typical school environments.

Belief in cognitive modifiability and human development. A teacher who is effective in diagnostic and formative classroom assessment should believe in cognitive modifiability and the potential for student development on a continuum, and in the utility of dynamic, formative assessment techniques in modifying levels of cognitive functioning in all learners. Human capacities, including intellectual, social, and emotional behaviors in any domain are now widely acknowledged to be both developmental and modifiable (Feuerstein, Rand, & Hoffman, 1979; Goleman, 1994; 2007; Kuhn, Katz, & Dean, 2004; Pressley, 1995; Siegler, 2000; Vygotsky, 1978; Wellman & Gelman; 1992).

Belief in proximal assessment, timely mediation, and learner development. The second indicator of effective diagnostic classroom assessment is a *Belief* in the importance of timeliness and proximity of interventions with learners, immediately following assessment (Feuerstein et al, 1979). Vygotsky (1978, p. 84) also stressed the value of proximal mediation during pedagogy, urging that teachers take advantage of a learner's "zone of proximal development" while teaching. Vygotsky pointed out that student performance should be assessed close-up, under the vigilant guidance of a more capable expert, who could be either a teacher or a peer with higher levels of competence in the domain.

Belief in the relationships among prior knowledge, scaffolding, situated learning, self-regulation, and development. Teachers who are effective in diagnostic assessment should also believe in principles of cognitive and educational psychology that are pertinent to optimizing student learning in specific domains. Some key principles point to the need for: (a) connecting new learning with prior knowledge levels of learners; (b) scaffolding instruction to build mental bridges and fill gaps in understanding; (c) situating learning in real life applications,; (d) understanding that learners draw on multiple concepts and a range of skills when attempting more complex tasks; (e) facilitating metacognitive skills in learners to help them regulate their own learning; and (f) providing risk-free and supportive environments to facilitate learner development in given areas (Anderson, 2005; Pellegrino et al, 2001).

Belief in the role of affect and regulated practice in human learning. When learners find value and meaning in tasks, they are motivated to persist, thereby reaching higher levels of cognitive competence (Kuhn, 2001). Talent development research in medicine, sports, and music shows that people who excel tend to dedicate hours of intrinsically motivated practice in a given area after their coaches point out their weaknesses. When such practice is highly concentrated and deliberate, it leads to greater levels of domain-related proficiency (Ericsson, 2004).

Table 1
The PALD Proficiency Domain for Teachers: Theoretically-derived Indicators of Teacher Belief

1.0 Believes in the utility of diagnostic classroom assessment for fostering student learning.	
General indicator(s)	Specific indicator(s)
1.1 Belief in cognitive modifiability and continuous human development in a domain	As a part of teaching in a domain, believes that: 1.1.1 Learning and development is a continuous process for all learners 1.1.2 Assessment and learner development are inter-related 1.1.3 Learner success and growth are facilitated with diagnostic assessment and mediation 1.1.4 Learning/development occurs in both cognitive and non-cognitive domains 1.1.5 Affective states of the learner (motivation, value, comfort) influence learning and development
1.2 Belief that proximal, timely assessment and mediation helps bridge learning gaps	To facilitate learner development in a domain, believes in the utility of: 1.2.1 Embedding self-designed assessments for diagnosis of learner needs 1.2.2 Probing and performing error analysis to identify learning gaps 1.2.3 Interacting, modeling and targeting immediate mediations/interventions to support learner growth 1.2.4 Connecting students' prior learning with new learning goals 1.2.5 Scaffolding while teaching or testing 1.2.6 Encouraging student self-monitoring and reflection 1.2.7 Situating learning experiences in real life 1.2.8 Giving students planned practice 1.2.9 Nurturing positive student affect 1.2.10 Creating a risk-free assessment environment 1.2.11 Taking graduated steps (from simple to complex) towards difficult goals
1.3 Belief in student-centered, outcome-driven pedagogy	To facilitate learner development, believes in the importance of: 1.3.1 Clarifying learning outcomes and instructional goals when designing instruction and assessment 1.3.2 Aligning learning outcomes/goals with instruction and assessment 1.3.3 Coaching, re-teaching, and giving feedback to students after assessment 1.3.4 Using formative assessment results to alter and improve teaching strategies 1.3.5 Gathering evidence of new learning via formative assessment before performing summative assessment
1.4 Belief in self-capacity to conduct diagnostic and formative assessment in typical school environments.	Expresses belief in own ability to: 1.4.1 Make changes towards learning-centered and diagnostic assessment practices 1.4.2. Balance external assessment and accountability demands with formative and diagnostic learner assessment 1.4.3 Balance the external or organizational assessment culture with best practices in classroom assessment 1.4.4 Manage ongoing, diagnostic assessment tasks (designing tests/items, scoring, analyzing errors, making formative decisions) as part of regular work routine

Error analysis, probing, learning need diagnosis, and human development. Diagnostic strategies, such as, asking probing questions of learners, are useful techniques for error detection and fostering learning on a continuum. The literature support for these indicators come from disparate traditions in educational measurement, cognitive psychology and pedagogy (Nichols, 1994; Tatsuoka, 1983; Webb, Franke, De, Chan, Freund, Shein, & Melkonian, 2009).

Belief in outcome-driven instruction, assessment, and learning. A major concept laced through the literature is the need for teacher *Beliefs* in learner-centered, supportive assessment environments while conducting formative assessment (Stiggins, 2002). Sub-domain 1.3 in Table 1 is also supported by the literature in instructional design and domain-referenced assessment, which repeatedly emphasizes the need for teachers to communicate clear educational targets and assessment criteria to students, for improving their own teaching processes and supporting learner growth (Chatterji, 2003; Dick, Carey, & Carey, 2005; Gagne, 1997; Nitko, 1989; Fuchs & Fuchs, 2003; Price & Nelson, 2007).

Belief in self-capacity to conduct student-centered diagnostic assessment in school environments. This indicator deals with a teacher's *self-efficacy* in conducting diagnostic, classroom assessment within the organizational structure of schools (Bandura, 1997). To be effective, teachers must be able to commit to the processes, despite factors related to the organizational structure, larger assessment culture, and outside accountability pressures or political challenges. That school organizations pose barriers to diagnostic, learner-centered assessment was documented at the start of this paper. The need for cohesive organizational support systems to help teachers change practices, has also been recognized in the teacher education, assessment, and educational policy literatures (see American Federation of Teachers et al, 1990; Borko, 1997; Brookhart, 2011; Johnson, Wallace & Thompson, 1999; Youngs, 2001).

Preliminary Item Pool: Teacher Beliefs

Table 2 presents a sample of illustrative items aligned with a few general and more specific indicators in the *Belief* domain. Readers should note that this is a beginning item pool for the design of a possible self-report survey. As suggested, teachers could respond on a 5-point Likert scale ranging from *Strongly Agree* to *Strongly Disagree*, with a *Don't know/Uncertain* option at mid-point. On a cluster of items defining a scale (domain or sub-domain), a summated, composite score would be derived from numerically-weighted responses of individual teachers. The composite score would denote a teacher's location on the *Belief* continuum for the *PALD Proficiency* construct (Crocker & Algina, 2006).

Table 2
A Sample of Items tapping Teacher Beliefs in Diagnostic Classroom Assessment

General and specific indicator(s)	Items for a self-report survey
1.1 Belief in cognitive modifiability and continuous human development in a domain 1.1.1 Learning and development in an area is continuous 1.1.5 Affective states of the learner (motivation, value, comfort) influence learning and development	1. All students can succeed and grow in any subject area (1.1.1). 2. How students feel about what they learn influences how much they learn (1.1.5).
1.2 Belief that proximal, timely assessment and mediation helps bridge learning gaps Believes in the importance of: 1.2.1 Embedding self-designed diagnostic assessments during teaching is important for facilitating learning 1.2.2 Performing error analysis and probing 1.2.3 Interacting, modeling and targeting immediate mediation/interventions	3. Teachers become more aware of student needs when they embed their own assessments during teaching (1.2.1). 4. If teachers examined student mistakes more closely, students would learn more (1.2.2).

Note. Matching indicator number in parenthesis.

Although the examples in Table 2 are all positively-oriented, individual items may also be negatively-oriented. In such cases, on a Likert scale, the *Strongly Agree* response option would have the lowest numeric rating (1) instead of the highest (5). Negatively-oriented items should be appropriately interspersed during the assembly of an instrument to control for potential faking, provision of socially-desirable responses, or a cavalier use of fixed-response sets by respondents.

Domain of Teacher Practices: Case Study Findings

The findings from a grounded theory analysis using the three data sources are presented in turn next, providing the larger context of the teacher's lesson. Verbatim data in the form of classroom interaction vignettes have been deleted due to space restrictions but are available from the author on request. The cumulative, thematic findings are tabulated in Tables 3-5 with samples of items in the right-hand column aligned with specific indicators. Recurring indicators are mentioned in the description of findings, but listed *only once* in Tables 3-5.

Findings from a demonstration-coaching observation of a PALD unit. This session was recorded during a 2-day visit when the teacher taught a review unit in long division during Year 1. Excerpted assessment tasks and a diagnostic scoring rubric used by the teacher with her class are shown in Figures 1-2. In this instance, these assessment artifacts were developed jointly by the teacher and assessment coach as a part of the first PALD demonstration unit.

Consistent with PALD training regimens designed using the literature just discussed, the assessment items in Figure 1 show a developmental order, leading up to a complex, applied problem linked to long division competencies subsumed within the larger domain. The student-friendly scoring rubric in Figure 2 is intentionally analytic so as to facilitate diagnosis, but also aligned with competencies in the domain. Themes extracted through the coding procedure applied to four sequenced teaching segments are shown in Table 3.

Tables 1 and 3 may be compared to examine overlaps in the findings using these two data sources. Cross-validated indicators include: the identification of clear learning targets and outcomes, use of culminating assessment tasks as a pre-test or practice test, use of ordered formative assessment tasks, use of rubrics to communicate expectations, embedding probing or conducting error analysis of student work during instruction, provision of mediation and practice, and repetition of formative assessment cycles.

Findings from a teaching observation of a non-PALD unit. Towards the end-of the grade 6 year (Year 2), *LA* was observed teaching a lesson in combinations on her own, requiring numeric and graphic representation of combinatorial problems applied to real life. She was now engaging in significantly more diagnostic practices that appeared spontaneous. There was no record of *LA's* use of any formal error analysis or repeated formative cycles during the 30-minute lesson observation. However, her probing was frequent and targeted, and she was able to fit in both individual and group mediation during the observed lesson.

At the same time, *LA* was found to engage in one indicator that was inconsistent with the philosophy and theory underlying PALD—she tended to use assessment to control student behaviors. It was a strategy she seemed to employ by habit, to hold the class' or particular student's attention, get students to comply with specific content-related expectations, and to obtain behavioral compliance by reminding students of the upcoming external examination on which she and the school would be held accountable for gains. Although there were several overlapping themes in the observation with those in Table 3, the recurring indicators are not repeated in Table 4. Table 4 shows only the new *Practice* indicators as "added" indicators to the overall domain.

Figure 1
Two Developmentally-Ordered, Diagnostic Assessment Tasks in Long Division Unit

Task examples	Task specifications and embedded competencies	Item Difficulty Level: % students answering correctly in Grade 5 (*n*=16)
Part 1 Look at the problem below to answer Questions 1-6. $5\overline{)175}$ 1. The problem is telling me to: a. Multiply 5 times 175 b. Find out how many groups of 5 are in 175 c. Divide 5 into 175 equal parts d. Add 5 and 175 e. Do something else (explain):	Question designed to probe application of long division concepts. Item situated in an item series. Learners/examinees are expected to restate what the problem is asking by using the structured-response item format. Scaffolding is provided via multiple answer options that aim to guide learners' thinking.	.75 (Easy Item)
Part 2 I began to do the following problem but know I have made some mistakes. Please find and mark my mistakes. Then, show me how I could fix them by re-doing the problem. H T O 610 9) 552 -54 12 9 3 -0 3R	Probe for checking concept understanding and procedural skills in long division. This is a situated task in that learners apply long division vocabulary, identify errors, check solutions, and explain answer, when given a problem that is incorrectly solved. Task uses an unscaffolded, open-ended item format.	.42 (Difficult Item)

Figure 2
Pre-Unit Assessment Task with Scoring Rubric

Task:
Coconut cookies come in packets of 12. I would like to give 1 cookie to each 5th grader at (your school) to celebrate the last day of school (in June). There are 219 5th graders at (school name) Elementary. How many packets of cookies should I buy so that every 5th grader gets at least 1 cookie? Will there be any cookies left over? Show your plan to solve the problem. Then set up the numbers and solve it. Explain all the parts of your answer, and show how you would check if your answer is correct. Show all your work.

Rubric: Checking What I Know in Long Division

I can:
-Say in my own words what the long division problem is asking me to do
-Read a story problem and set up the long division and/or <u>other</u> operations needed to solve it
-Identify the "dividend" and "divisor" in an long division problem
-After the long division problem is set up, know how to start the algorithm (find digit with the highest place value in the dividend)
-Recall and use multiplication facts correctly when doing long multiplication and long division
-Recall place value concepts and give place value labels for digits in whole and decimal numbers, when doing long multiplication and long division
-Follow the steps of the long division algorithm correctly*
-Repeat the long division algorithm until I get a remainder that cannot be further divided (grade 5)
OR
-Continue the long division algorithm so I can express remainder as a decimal of the quotient (grade 6)
-Check the long division answer with "backwards" operations
-Explain what the answer means in my own words
-Think-aloud and write the steps of long division
-Think aloud and explain the answer and parts of the answer in a story problem
-Keep my scratch work neat (so that I don't get lost)
-Look at my mistakes to help me grow (without fear)
-Not give up before finishing a problem
-Teach long division to a friend
-Make connections between fractions, decimals and division/ long division operations

** Students used a mnemonic in class—Divide (Dad), Multiply (Mom), Subtract (Sister), Bring Down (Brother), Again (Aunt), Remainder (Repeat? End): DMSBAR*

Table 3
PALD Proficiency Domain based on Case Study: Teacher Practice Indicators derived from Coaching Session

2.0 Demonstrates diagnostic classroom assessment practices while teaching	
General and specific indicator(s)	Matching items for a behavior-based observation tool or self report survey
2.1 Clarifies targeted goals, outcomes, and embedded competencies on a continuum to prepare for student need diagnosis 2.1.1 Identifies embedded learning outcomes, concepts and skills from a culminating target (problem or task) 2.1.2 Designs/uses ordered items or tasks that test mastery of embedded concepts and skills in a unit 2.1.3 Gauges difficulty of goals vis-à-vis student background levels 2.1.4 Designs/uses pre-tests, homework, and class work exercises for formative assessment of student learning needs 2.1.5 Evaluates difficulty of assessment tasks vis-à-vis student levels and learning targets	1. I use more difficult problems in my early unit assessments to identify student needs (2.1.2). 2. My goals for a unit are broken down to match levels of different students (2.1.4). 3.. I give homework exercises to check where students are in learning new goals for a unit (2.1.5).
2.2 Plans and delivers instruction to class linked to formative assessment results 2.2.1 Deliberately plans instruction using assessment results from pre-tests, mid-unit homework, or classwork exercises before using assessment results for grading 2.2.2 Makes time to examine student work formatively 2.2.3 Identifies student errors/mistakes vis-à-vis embedded competencies and learning targets 2.2.4 Identifies mediation techniques specific to observed learning gaps in students 2.2.5 Tries more than one mediation strategy with students before moving on 2.2.6 Repeats formative assessment and teaching cycles, as needed (re-teaches and re-assesses) 2.2.7 Makes time to summarize and review student responses to tests/assessments (analyzes data quantitatively or qualitatively) 2.2.8 Encourages student engagement or dialogue about concepts and skills relevant to domain/unit mastery	4. I use the pre-unit assessments results to plan my lessons. (2.2.1) 5. I test formatively to see if students make errors in the concepts I am teaching (2.2.2). 6. Depending on errors my students make on exercises I give them, I change what I teach (2.2.4). 7. I conduct formative assessment more than once in a unit (2.2.6).

Note. Matching indicator number in parenthesis.

Table 4

PALD Proficiency Domain: Added Teacher Practice Indicators derived from Classroom Observations

2.0 Demonstrates diagnostic classroom assessment practices while teaching	
General and specific indicator(s)	Matching items for a behavior-based observation tool or self report survey
2.2 Plans and delivers instruction linked to formative assessment results 2.2.9 Uses situated tasks meaningful to learners during assessment and instruction. 2.2.10 Probes student thinking about mistakes related to concepts in unit 2.2.11 Encourages examination of errors by students 2.2.12 Mediates individually while teaching 2.2.13 Mediates with whole class or small groups while teaching 2.2.14 Creates a supportive, learning-oriented assessment culture	8. I use everyday examples during formative assessment and instruction (2.2.9). 9. I ask students to think about their mistakes (2.2.11). 10. I re-teach concepts to the whole class, when needed (2.2.13).
Negatively-oriented indicator 2.2.15 Uses classroom assessment to force behavioral compliance from students or for behavior management	Negatively-oriented items 11. If they are not attentive when I teach, I warn students of external exams.

Note. Matching indicator number in parenthesis.

Findings from end-of project teacher report. *LA*'s reflective report, where she indicated having applied PALD with a unit on decimals and fractions, involved an extension of the long division unit. In it, she regretted that "*…due to time constraints, it was impossible for me to intervene (with) all students consistently,* ". However, several recurrent themes were now evident that were consistent with PALD theory.

The themes generated through analysis of these records are reported in Table 5. As shown, a new negative indicator surfaced with regard to student attitudes and behaviors. Specifically, *LA* reported giving up with four students, communicating negative *Beliefs* about their capacity to master the concepts and engage with the material in a meaningful way. Regardless, she observed that the benefits of the PALD experience were largely affective, and that "*students came to expect a lot from the assessments*".

Table 5
PALD Proficiency Domain: Added Teacher Practice Indicators derived from Teacher Journal Records

2.0 Demonstrates diagnostic classroom assessment practices while teaching	
General and specific indicator(s)	Matching items for a behavior-based observation tool or self report survey
2.2 Plans and delivers instruction to class linked to formative assessment results 2.2.16 Uses formative assessment tasks in a variety of formats 2.2.17 Repeats error analyses of student work, as needed 2.2.18 Tracks student growth via formative assessment records, quantitatively or qualitatively 2.2.19 Communicates performance expectations to students via rubrics or other means 2.2.20 Rewards/recognizes students for participating in formative assessment processes 2.2.21 Provides practice exercises as part of formative assessment cycle	12. I use different types of assessment tasks—written, oral, projects–during instruction (2.2.16). 13. I keep records of student growth on particular skills (2.2.18). 14. I recognize students publicly in class if they show growth after a formative assessment. (2.2.20).
Negatively-oriented indicator 2.2.22 Communicates negative beliefs about students' learning capacities	Negatively-oriented items 15. I let weak students know they may not master the entire unit.

Note. Matching indicator number in parenthesis.

Discussion

In general, the *Belief* and *Practice* sub-domains of *PALD Proficiency* were cross-validated using the literature review and data from the teacher's classroom (Questions 1-2). A good deal of consistency is evident in the operational definitions presented in Tables 1-2 and Tables 3-5. Recurrent themes were evident in later observations and journal records that suggested triangulation across data sources. Since instruments have yet to be assembled and the item pool further validated, it may be premature to replace one set of indicators with the other.

In terms of a "theory of practice", the case study data suggested a continuum of teacher change that may be useful in setting realistic expectations for teacher in-service training (Question 3). A veteran teacher like *LA* could make positive shifts towards diagnostic assessment-related attitudes and behaviors over the course of two years. In the teacher's case, several PALD practices showed transfer from the scaffolded delivery of the long division PALD unit, to other units that the teacher taught on her own. Affective changes were also recorded for both students and the teacher with regard to assessment. However, the depth of teacher skills attained may need continuing attention. The selection of particular content-related intervention and mediation strategies by the teacher may have been restricted to her comfort zones (e. g., *LA* frequently used peer tutoring, and her mediation techniques were more varied with particular concepts, like place value).

The barriers to teacher change in this case were reflected in some fixed mind-sets about the capacity of a few students to learn, a tendency to use classroom assessment to manage and control student behaviors, and rationalization of the press related to accountability testing as a motivator to push for student achievement. As suggested in the literature, while beliefs tap into a seemingly different dimension of attitude, behaviors and practices are likely to be alternate manifestations of the same underlying construct. These issues may be examined empirically through future research. Teacher training efforts could focus on the noted barriers when implementing classroom assessment reforms.

It should also be noted that the same general indicator framework for teacher *Beliefs* could be re-formulated to tap *Practice* indicators. Although empirical correlations stand to be verified, dual configurations and interpretations of the same indicator in an attitudinal-behavioral construct would be consistent with the tripartite model of measurement cited earlier. As illustrated in Tables 2-4, alternate scale configurations may be accomplished by changing the syntax and language structure to incorporate appropriate verbs in items, and linking the items with matching response scales.

As the literature surveyed indicates, the lack of adequate formative and diagnostic assessment in scholastic settings is an international concern. Citations in this paper are from observers and researchers in the U.K., U.S., and other nations. In this sense, this study fills a practical void documented across these regions. The cross-validated domain of *PALD Proficiency* indicators may therefore prove useful for designing formal tools to assess, coach, or guide teacher practices in future professional development or research programs. In terms of new directions, the study permitted a demonstration of a mixed methods approach to domain specification and instrument design. The utility of this methodological approach should continue to be tested by future researchers.

To conclude, limitations of the present effort must be borne in mind. First, the case data here are based on a single classroom with limited generalizability. As new literature and empirical studies emerge, content-based validation of the domain and items should continue. Further research should also continue with multiple teacher cases to replicate and broaden the findings. Second, the sample of items is not an exhaustive list and should be expanded. Lastly, items and eventual instruments developed using the framework should undergo appropriate psychometric evaluations before formal applications are undertaken on a larger scale. Efforts to support teacher development in diagnostic classroom assessment should continue.

References

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

American Federation of Teachers, National Council on Measurement in Education, and National Education Association. (1990). Standards for teachers competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9*(4), 30-32.

Andeson, R. (2005). *Cognitive psychology and its implications* (6th ed.). New York, NY: Worth Publishers.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice.* Buckingham: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.

Borko, H. (1997). New forms of classroom assessment: Implications for staff development. *Theory into Practice, 36*, 231-238.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3-12.

Chatterji, M. (2003). *Designing and using tools for educational assessment.* Boston, MA: Allyn & Bacon.

Chatterji, M., Koh, N., Choi, L., & Iyengar, R. (2009). Closing learning gaps proximally with teacher-mediated diagnostic classroom assessment. *Research in the Schools*, *16*(2), 59-75.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed method approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory.* Wadsworth Publishing Company.

Dick, W., & Carey, L. (2005). *The systematic design of instruction* (6th ed.). Boston, MA: Allyn and Bacon.

Erickson, F. (2007). Some thoughts on "proximal" formative assessment of student learning. In P. A. Moss (Ed.), *Evidence and decision making* (pp. 186-216). Malden, MA: Blackwell.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*, S70-S81.

Feuerstein, R., Rand, Y., & Hoffman, M. (1979). *The dynamic assessment of retarded performers: The learning assessment potential device, theory, instruments and techniques.* Baltimore, MD: University Park Press.

Fink, A. (1995). *How to design surveys.* Thousand Oaks, CA: Sage Publications.

Fuchs, L. S., & Fuchs, D. (2003). Can diagnostic reading assessments enhance general educators' instructional differentiation and student learning? In B. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 325-351). Baltimore, MD: York Press.

Gagne, R. (1997). Mastery learning and instructional design. *Performance Improvement Quarterly, 10*(1), 8-19.

Goleman, D. (1994). *Emotional intelligence: Why it can matter more than IQ.* New York, NY: Bantam.

Goleman, D. (2007). *Social intelligence: The new science of human relationships.* New York, NY: Bantam-Dell.

Greene, J. C., & Caracelli, V. J. (Eds.). (1997). *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms.* New Directions in Evaluation. San Francisco, CA: Jossey-Bass.

Johnson, S. T., Wallace, M. B., & Thompson, S. D. (1999). Broadening the scope of assessment in the schools: Building teacher efficacy in student assessment. *The Journal of Negro Education, 68*(3), 397-408.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kuhn, D. (2001). How do people know? *Psychological Science, 12*(1), 1-8.

Kuhn, D., Katz, J. B., & Dean, D. (2004). Developing reason. *Thinking & Reasoning, 10*(2), 197–219.

Lave, J., & E. Wenger. (1991). Situated learning: Legitimate peripheral participation. Cambridge: Cambridge University Press.

Mabry, L. (2009). Case study methods in educational evaluation. In K.E. Ryan & J.B. Cousins (Eds.), *The SAGE international handbook of educational evaluation* (pp. 341-356). Thousand Oaks, CA: Sage.

Macintyre, J., Latta, M., Buck, G., & Beckenhauer, A. (2007). Formative assessment requires artistic vision. *International Journal of Education & the Arts, 8*(4), 23-36.

Nichols, P. D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 447-474). New York: American Council on Education and MacMillan Publishing Co.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & (Eds.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academies Press.

Popham, W. J. (2008). *Transformative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Pressley, M. (1995). What is intellectual development about in the 1990s? Good information processing. In F. Weinert W. Schneider (Eds.), *Memory performance and competencies: Issues in growth and development.* Hillsdale, NJ: Erlbaum.

Price, K. M., & Nelson, K. L. (2007). *Planning effective instruction: Diversity responsive methods and management (3rd ed.).* Belmont, CA: Thomson Wadsworth.

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.

Rosenberg, M. J., & Hovland, C. I. (1960). Cognitive, affective, and behavioral components of attitude. In M. J. Rosenberg, C. I. Hovland, W. J. McGuire, R. P. Abelson, & J. W. Brehm (Eds.), *Attitude organization and change: An analysis of consistency among attitude components* (pp. 1-14). New Haven, CT: Yale University Press.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger.

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

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 623-646). Westport, CT: Praeger.

Siegler, R. (2000). The rebirth of children's learning. *Child Development, 71*, 26–35.

Stake, R. E. (1997). Case study methods. In R. M. Jaeger (Ed.), *Complementary methods for research in education* (pp. 401–421). Washington, DC: American Educational Research Association.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Strauss, A. C., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Webb, N. M., Franke, M. L., De, T., Chan, A. G., Freund, D., Shein, P., & Melkonian, D. K. (2009). 'Explain to your partner': Teachers' instructional practices and students' dialogue in small groups. *Cambridge Journal of Education, 39*(1), 49–70.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43*, 337-375.

Youngs, P. (2001). District and state policy influences on professional development and school capacity. *Educational Policy, 15*(2), 278-301.

Appendix

Coding running observation records using grounded theory analysis: How indicators were derived from a teacher-class interaction during a unit on probability

Theme/ Indicator	Description	Running record	Sentence by sentence codes applied to text segments
Probing	Uses probing questions to check for concept understanding, mediates with whole class, and demonstrates as a part of instructional exchange specific to a math problem/concept.	Teacher uses a bag of letters (Scrabble pieces) to teach probability concepts and vocabulary related to dependent and independent observations. T: Well, dependent means it relies on something else. ***And I put the M back. Right? And I reach in the bag and pull another letter. Do those pulls affect each other?*** S (2): No	(T uses materials to teach probability-related vocabulary) (T clarifies word 'dependent') (T probes with 3 Qs while demonstrating) (S gives content-related response)
Probing		T: ***You think it's gonna change the probability on one?*** Ss (1): Yes, yes. Ss (2): No, no.	(T probing) (S gives content-related response) (S gives content-related response)
Re-teaching, demonstrating concept		T (repeats): ***I reach into the bag and pull a letter out, look at it, put it back in, and reach in the bag and pull another one out.*** Ss: (Unanimously) No	(T clarifies and demos again) (S content-related response)
Giving feedback-- large group mediation.		T: They're independent. One does not affect the other. That's independent. But the other example that Sean picked up on, is that if I pulled out the M and I put it over on the side, and I reach in the bag again and pulled out another letter. ***Do you understand how the second time, my probability is not out of 11 anymore?*** Ss (most of class): Yeah	(T Feedback during direct concept instruction) (T repeats concept, vocabulary, and demo) (T probes with 1 Q.) (Ss' give content-related response)
Probing. Giving feedback (reinforcing), large group mediation.		T: So my first event had an effect on my second event. Okay. ***Do you understand that? That means they're dependent events. Okay?*** That's some review on our vocabulary. Any questions on the problem?	(T Feedback during direct concept instruction)

About the Author

Madhabi Chatterji is Associate Professor of Measurement, Evaluation, and Education at Teachers College, Columbia University where she directs the Assessment and Evaluation Research Initiative (AERI), a center dedicated to promoting valid and meaningful use of educational assessment information, nationally and globally. Her research and teaching interests include instrument design and validation, mixed methods, evidence standards and evidence synthesis methods, standards-based reforms, and educational equity. She was the Principal Investigator of the reported study, funded by the National Science Foundation.