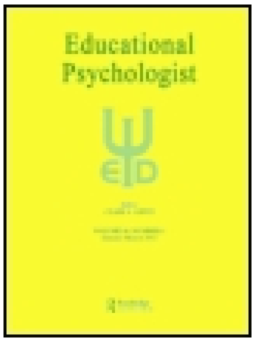


Routledge
Taylor & Francis Group

Educational Psychologist

ISSN: 0046-1520 (Print) 1532-6985 (Online) Journal homepage: http://www.tandfonline.com/loi/hedp20

A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments

James W. Pellegrino, Louis V. DiBello & Susan R. Goldman

To cite this article: James W. Pellegrino, Louis V. DiBello & Susan R. Goldman (2016): A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments, Educational Psychologist, DOI: 10.1080/00461520.2016.1145550

To link to this article: http://dx.doi.org/10.1080/00461520.2016.1145550

Published online: 09 Mar 2016.

Submit your article to this journal

View related articles

View Crossmark data

EDUCATIONAL PSYCHOLOGIST, 0(0), 1–23, 2016
Copyright © Division 15, American Psychological Association
ISSN: 0046-1520 print / 1532-6985 online
DOI: 10.1080/00461520.2016.1145550

Routledge
Taylor & Francis Group

A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments

James W. Pellegrino, Louis V. DiBello, and Susan R. Goldman
Learning Sciences Research Institute
University of Illinois at Chicago

Assessments that function close to classroom teaching and learning can play a powerful role in fostering academic achievement. Unfortunately, however, relatively little attention has been given to discussion of the design and validation of such assessments. The present article presents a framework for conceptualizing and organizing the multiple components of validity applicable to assessments intended for use in the classroom to support ongoing processes of teaching and learning. The conceptual framework builds on existing validity concepts and focuses attention on three components: *cognitive validity, instructional validity*, and *inferential* validity. The goal in presenting the framework is to clarify the concept of validity, including key components of the interpretive argument, while considering the types and forms of evidence needed to construct a validity argument for classroom assessments. The framework's utility is illustrated by presenting an application to the analysis of the validity of assessments embedded within an elementary mathematics curriculum.

Over the last 15 years, numerous developments in the fields of cognitive psychology, educational psychology, the learning sciences, and educational measurement have begun to reshape the field of educational assessment. Among these is the realization that assessment is fundamentally a process of reasoning from evidence that needs to be guided by theories, models, and data on the nature of knowledge representations and the development of competence and expertise in typical domains of classroom instruction (e.g., Pellegrino, Chudowsky, & Glaser, 2001). Other key understandings of assessment and its design and validation include the multiple purposes assessment may serve in the educational system, including formative, summative, and evaluative functions, and that the design of any assessment must be optimized for its intended purpose and interpretive use (e.g., Black & Wiliam, 1998; Gordon Commission on the Future of Assessment in Education, 2013a, 2013b; Mislevy & Haertel, 2006; Pellegrino et al., 2001; Wiliam, 2012). Coinciding with these developments is contemporary thought that the validity of a given assessment should be construed as an argument consisting of claims related to an assessment's intended interpretive use and the forms of evidence that provide the warrants in support of those claims (e.g., Kane, 2006; 2013). One example would be the claim that a computer-based assessment can be used for diagnosis of a students' conceptual understanding of linear functions for instructional planning. Part of the evidence for this claim would be reliable differentiation of student performance in terms of instructionally meaningful patterns of correct and incorrect answers, where these patterns are interpretable as indicators of what students understand and misunderstand about important properties of functions. Another example would be the claim that a standardized reading assessment can be used to measure students' proficiency in comprehending expository text for purposes of predicting college readiness in English language arts. Part of the evidence would be the magnitude of the correlation

Correspondence should be addressed to James W. Pellegrino, Learning Sciences Research Institute, University of Illinois at Chicago, 1240 West Harrison (M/C 057), Chicago, IL 60607. E-mail: pellegjw@uic.edu

between the test scores and freshman grade point average in English Language Arts or other courses.

Much of the literature on assessment validity has concentrated on assessments designed for large-scale and often high-stakes uses in the educational system, many of which are associated with various policies related to accountability. The reading comprehension example just mentioned is illustrative of such cases. Many such assessments have limited conceptual bases with respect to contemporary cognitive models and theories about the nature of knowledge and the acquisition of competence, although that is changing slowly (e.g., Partnership for Assessment of Readiness for College and Careers, 2014; Smarter Balanced Assessment Consortium, 2014). In addition, increasing attention is being paid to the need for systems of assessments, with an emphasis on the design and use of assessments intended to function much closer to the processes of teaching and learning (e.g., National Research Council [NRC], 2003; Pellegrino et al., 2001; Pellegrino, Wilson, Koenig, & Beatty, 2014; Wilson & Bertenthal, 2006). Thus, there is a need for careful consideration regarding the design and validation of assessments intended for classroom use for both formative and summative purposes. Many such assessments are tied more closely to detailed analyses of the instructional domain (e.g., reading, mathematics, science, history) and theories, models and data regarding student knowledge representations, and how those change over time as a consequence of instruction. Examples include work on learning progressions and learning trajectories in areas such as mathematics and science (e.g., Alonzo & Gotwals, 2012; Corcoran, Mosher, & Rogat, 2009; Daro, Mosher, Corcoran, & Barrett, 2011).

Despite the significant increase in efforts to frame the focus of discussions of assessment in terms of domain-based theories and models of cognition and learning, and an increasing attention to the use of assessment in the classroom to support various teaching and learning functions, there has been a paucity of discussion about the meaning of validity for assessments intended to function close to instruction. Such a discussion should include consideration of the contributions of cognitive and psychometric theories and methods in the design of contemporary assessments and in the collection of data that contribute to the process of validation. As framed in the language of Kane (2006, 2013) and others, this involves constructing the interpretive argument and marshaling evidence in support of that argument. The present article provides just such a discussion of validity by presenting a framework for conceptualizing the multiple components of validity applicable to assessments intended to function at the classroom level to support ongoing processes of teaching and learning. Our goal is to clarify the concept of validity, including key components of the interpretive argument together with the types and forms of evidence needed to construct a validity argument for such assessments. We also illustrate the value of this framework by presenting an application to the analysis of assessments embedded within an elementary mathematics curriculum.

In the first section we discuss assessment validity in terms of the logic of argument and evidence. We propose a framework that identifies salient components of validity for instructional settings in which assessments can be used to provide direct benefits to both instructors and students. Our framework builds on existing validity ideas and encompasses three components: *cognitive validity, instructional validity*, and *inferential* validity. Evidence for each component can come in multiple forms, and it is associated with aspects of the design of an assessment, characteristics of the tasks and performances that make up the assessment, and aspects of expected and actual student performance on those tasks.

In the second section we consider fundamental principles for the conceptualization and design of assessments. Such principles are especially relevant and applicable to constructing validity arguments for assessments intended to function close to the processes of teaching and learning. We begin by discussing the logic of assessment in terms of the *assessment triangle*. The latter conceptualizes assessment as a process of evidentiary reasoning involving the three critical and interconnected elements of cognition, observation, and interpretation (Pellegrino et al., 2001). We then consider the role of theories and data about student cognition and learning in guiding and constraining assessment development. Because assessment development, as guided by theory and data, is dependent on translation through processes of design, we then discuss principled assessment design practices. Two such approaches are evidence-centered design (ECD) (Mislevy & Riconscente, 2006), and the construct-modeling approach (Wilson, 2005; Wilson & Draney, 2004; Wilson & Sloane, 2000). Both approaches involve identifying the claims to be made about student knowledge and how those claims are connected to the specification and collection of multiple forms of evidence that would lend support to the claims.

Although assessment design processes are key elements in arguments and evidence about validity, different forms of empirical evidence are also needed to determine how well an assessment actually performs relative to claims about what the assessment was designed to do. In the third section we consider multiple forms of evidence that are relevant to support claims about any given assessment's validity. The sources of evidence are related to each of the three components of validity discussed in the first section. In the fourth section we illustrate application of the conceptual and evidence framework to a set of assessments embedded within a well-known and commonly used elementary grades mathematics curriculum *Everyday Mathematics* K-6 (EDM) (Bell & Bell, 2007). The article concludes with a consideration of additional issues regarding application of the proposed validity analysis framework, including the need for careful and consistent application of this logic in the design of integrated curriculum, instruction, and assessment resources. Such resources are critical to support teachers and students in attaining the deeper learning goals intended by the Common Core Standards in Mathematics and English Language Arts (2010a, 2010b) and the Next Generation Science Standards (Achieve, 2013).

ARGUMENTATION, EVIDENCE, AND MULTIPLE COMPONENTS OF VALIDITY

The joint American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA/APA/NCME; 1999/2014) frame validity largely in terms of "the concept or characteristic that a test is designed to measure" (p. 5). In Messick's construct centered view of validity, the theoretical construct the test score is purported to represent is the foundation for interpreting the validity of any given assessment (Messick, 1994). For Messick (1989), validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13). Important work has been done to refine and advance views of validity in educational measurement (see, e.g., Haertel & Lorie, 2004; Kane, 1992, 2001, 2002, 2006, 2013; Mislevy, 1996; Mislevy, Steinberg, & Almond, 2003). Contemporary perspectives call for an interpretive validity argument, an argument that "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p. 23).

Kane (2006) and others (Haertel & Lorie, 2004; Mislevy et al., 2003), distinguish between (a) the interpretive argument, that is, the propositions that underpin test score interpretation, and (b) the evidence and arguments that provide the necessary warrants for the propositions or claims of the interpretive argument. In essence this view identifies as the two essential components of a validity argument the claims being made about the focus of an assessment and how the results can be used (interpretive argument), together with the evidence and arguments in support of those claims. Returning to the example used earlier about linear functions, the claim would be that a given computer-based assessment, given the content and structure of its tasks, can be used for diagnosis of students' conceptual understanding of linear functions for instructional planning. Part of the evidence for this claim would be reliable identification of patterns of correct and incorrect answers that are interpretable as indicators of student understanding and misunderstanding of important properties of linear functions.

Mislevy et al. (2003) asserted that validity evidence can be represented formally in terms of a structured argument, an approach initially introduced and outlined in the context of legal reasoning by Toulmin (2003). In Toulmin's view, a good argument is one that provides strong evidentiary support for a claim that, in turn, permits it to withstand criticism. Toulmin's practical approach begins by explicating the various "claims of interest" and then provides justification for those claims by identifying the evidence (data and/or expert opinion) to support those claims. The approach also calls for supplying a "warrant" that interprets the data and explicates how the data support the claims of interest. Toulmin's work, historically, focused on the formal role of argumentation and was used to evaluate the rational basis of arguments presented typically in the courtroom. Appropriating this approach, contemporary educational measurement theorists have framed test validity as a reasoned argument backed by evidence (e.g., Kane, 2006). An argument and evidence framing of validity supports investigations for a broad scope of assessment designs and purposes, including many that go beyond typical large-scale tests of academic achievement or aptitude. An example of such a novel application is the framework presented here for assessments intended for classroom use.

For classroom assessment contexts and purposes, we propose a specific validity framework that takes into account the various forms of validity that have typically been discussed for large-scale, standardized tests (e.g., AERA, APA, NCME, 1999/2014; Messick, 1989). The three forms of validity discussed next build on current thinking that an assessment is not valid or invalid per se (i.e., validity is not a property of the instrument per se), but rather that validity has to be judged relative to the intended interpretive use of the results. Furthermore, for a given intended interpretive use, multiple aspects of validity can be evaluated, each of which deserves attention in its own right. These multiple aspects include what has traditionally been termed *content validity* (the extent to which an assessment represents critical aspects of the domain being assessed), *criterion validity* (how well assessment scores correlate with external criterion measures of interest), *construct validity* (determining what really is being assessed and how well that corresponds to what was intended to be assessed), and *consequential validity* (the use of assessments and their positive and negative consequences for teaching and learning).

Our approach to conceptualizing the validity of instructionally relevant assessments is not at variance with current discussions of validity such as those reflected in the standards for educational and psychological testing (AERA/APA/NCME, 1999/2014). Rather, we argue that our framework builds on and reflects multiple and complementary facets of validity as articulated in those standards, but it organizes and focuses them for the intended interpretive uses of classroom assessments. In particular, we are focused on those aspects of knowledge and skill that are the targets of assessment within instructional settings and the quality and value of the information about those constructs relative to support of ongoing classroom teaching and learning. Our three components of validity, their rationales, and what they are designed to address are as follows:

1. **Cognitive**. This component addresses the extent to which an assessment taps important forms of domain knowledge and skill in ways that are not confounded with other aspects of cognition such as language or working memory load (the construct). Cognitive validity should be based on what is known about the nature of student cognition and understanding in areas of the curriculum such as literacy, mathematics, and science and how it develops over time with instruction to determine what knowledge and skills students are supposed to use and those that they actually do use when interacting with the assessment (e.g., Bransford, Brown, Cocking, Donovan, & Pellegrino, 2000; Duschl, Schweingruber, & Shouse, 2007; Kilpatrick et al., 2001).
2. **Instructional**. This component addresses the extent to which an assessment is aligned with curriculum and instruction, including students' opportunities to learn, as well as how it supports teaching practice by providing valuable and timely instruction-related information. Instructional validity should be based on evidence about the alignment of the assessment with the knowledge and skills of interest as defined by standards and curricula as well as the practicality and usefulness of the instrument and the assessment outcome information for teachers and students as a guide to instruction and learning (e.g., Black, Harrison, Lee, Marshall, & Wiliam, 2004; Heritage, 2010; Kingston & Nash, 2011; Wiliam, 2007).
3. **Inferential**. This component is concerned with the extent to which an assessment reliably and accurately yields model-based information about student performance, especially for diagnostic purposes. As such, it is very closely connected with claims associated with the cognitive and instructional validity of an assessment. Inferential validity should be based on evidence derived from various analytic methods, including multivariate measurement and statistical inference (e.g., DiBello, Roussos, & Stout, 2007; van der Linden & Hambleton, 1997), to determine whether task performance reliably aligns with one or more underlying conceptual measurement models that are appropriate to the intended interpretive use.

These three components of validity are identified as most salient for assessments intended for use in the classroom close to instruction and represent a repurposing of the standard validity components. For instance, cognitive validity includes elements of Messick's construct validity, instructional validity incorporates traditional content and consequential validity, and inferential validity is related to criterion validity. By contrast with the earlier collection of components, the coordinated framework discussed here provides a coherent basis for the collection, organization, and interpretation of validity evidence for the class of assessments intended to support classroom teaching and learning. This validity framework constitutes an example of an integrated argument and evidence framework for this class of assessments.

The three components, individually and collectively, are critical for evaluating the validity of assessments that are intended to support learning and instruction. For example, an assessment may adequately reflect key aspects of a cognitive construct such as the knowledge and skill associated with performing mathematical operations on fractions but be structured in such a way that it provides little in the way of reliable diagnostic information that informs a teacher as to the particular facets of knowledge that her students have mastered versus those that are problematic. In other words, it might have considerable cognitive validity but have very limited instructional and/or inferential validity. Thus, in evaluating the validity of assessments intended to function at the classroom level for formative or summative purposes, one needs to collect evidence to support each of the three components of validity. Such evidence can come from multiple sources that are complementary and convergent (see the Multiple Forms of Evidence section for a more extensive discussion of multiple forms of evidence).

The compilation of evidence and construction of a *validity argument* for a given assessment should be an ongoing activity that begins during assessment design and continues through various iterations from pilot testing of the assessment materials and procedures to subsequent operational versions that might be used on various scales of implementation—classroom, school, district, state, and/or national. Thus, the three components in our proposed framework can serve as a guide to determining the validity of an assessment both prospectively, that is, during its conceptualization and design, and retrospectively, that is, for purposes of evaluation of the relative strengths and weaknesses of any given assessment that is being used by educators proximal to the processes of teaching and learning.

ASSESSMENT OF STUDENT KNOWLEDGE: CONCEPTUALIZATION AND DESIGN

The validity of an assessment very much depends on the conceptual frameworks and processes that have been used to guide development of the assessment given its intended use. As argued in this section, these frameworks and processes constitute a key part of the argument and evidence about an assessment's cognitive, instructional, and inferential validity.

Assessment as a Process of Evidentiary Reasoning: The Assessment Triangle

Educators assess students to learn about what they know and can do, but assessments do not offer a direct pipeline into a student's mind. Assessing educational outcomes is not as straightforward as measuring height or weight; the attributes to be measured are mental representations and processes that are not outwardly visible. Thus, an assessment is a tool designed to observe students' behavior and produce data that can be used to draw reasonable inferences about what students know. Deciding what to assess and how to do so is not as simple as it might appear.

The process of collecting evidence to support inferences about what students know represents a chain of reasoning about student learning from evidence that characterizes all assessments, from classroom quizzes and standardized achievement tests, to computerized tutoring programs, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. In the 2001 report *Knowing What Students Know: The Science and Design of Educational Assessment*, issued by the NRC, the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino et al., 2001). The vertices of the assessment triangle (see Figure 1) represent the three key elements underlying any assessment: a model of student *cognition* and learning in the domain of the assessment, a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies relative to the cognitive model, and an *interpretation* process for making sense of the evidence in light of the assessment purpose and student understanding. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, or evaluated, without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the *Knowing What Students Know* report is that for an assessment to be effective and valid, the three elements must be in synchrony. The assessment triangle provides a useful framework for analyzing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing their validity (e.g., see Marion & Pellegrino, 2006).

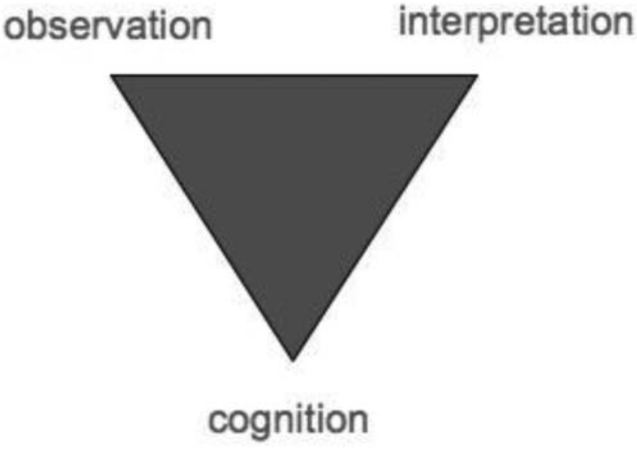


FIGURE 1. The Assessment Triangle.

The *cognition* corner of the triangle refers to theory, models, data, and a set of assumptions about how students represent knowledge and develop competence in a subject matter domain (e.g., fractions; Newton's laws; thermodynamics). In any particular assessment application, a theory of learning in the domain is needed to identify the set of knowledge and skills that is important to measure for the intended context of use, whether that be to characterize the competencies students have acquired at some point in time to make a summative judgment or to make formative judgments to guide subsequent instruction so as to optimize future learning. A central premise is that the cognitive theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in the domain being assessed. The theoretical propositions must be expressed at a grain size or level of detail appropriate to the assessment's intended use. Thus, this aspect of the assessment triangle forms the foundation for arguments related to an assessment's cognitive validity as well as other aspects of validity that logically follow from it. For an instructional target such as knowledge of fractions, the cognitive model could include the conceptual and procedural knowledge that would allow a student to express the magnitude of a fraction as a real number via mapping the value onto a number line or compare the magnitude of two fractions, determine which was larger, and express their values as real numbers.

Every assessment is also based on a set of assumptions and principles about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills reflected in the cognitive model. The tasks to which students are asked to respond on an assessment are not arbitrary. They must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be made on the basis of the assessment results. In the example just given about fractions, the tasks must be designed to reveal the desired forms of knowledge and understanding. Thus, they are linked to arguments about cognitive validity as well as instructional and inferential validity. They must be appropriate given the context of instruction and students' prior opportunity to learn, and they must provide evidence that can be reliably interpreted relative to the underlying cognitive constructs.

The *observation* vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students for the assessment's purpose. In assessment, one has the opportunity to structure some small corner of the world to make observations. The assessment designer can use this capability to maximize the value of the data collected, as seen through the lens of the underlying assumptions about how students learn in the domain, what is important for them to know and be able to do, and what students would be expected to know given their instructional history and opportunity to learn.

Finally, every assessment embodies certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. As such it characterizes the inferential validity of the assessment in terms of how the observations are to be used in providing information relative to the cognitive model and with respect to any possible instructional implications.

In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterization or summarization of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher and is usually based on an intuitive or qualitative model rather than a formal statistical one. Even informally, teachers make coordinated judgments about what aspects of students' understanding and learning are relevant, how a student has performed on one or more tasks, and what the performances mean about the student's knowledge and understanding. Their ability to make such a judgment is tied to the quality and interpretability of the information available from the assessments they are using.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Collectively, the cognition, observation, and interpretation elements establish key aspects of the argument about an assessment's cognitive, instructional, and inferential validity. Thus, to have a valid and effective assessment, all three vertices of the triangle must work together in synchrony. Central to this entire process are theories and data on how students learn and what students know as they develop competence for important aspects of the curriculum.

The Role of Models of Domain-Specific Learning: Learning Progressions

As just argued, the targets of inference for any given assessment, that is, "what knowledge and skills are being reasoned about," should be largely determined by models of cognition and learning that describe how people represent knowledge and develop competence in the domain of interest (the *cognition* element of the assessment triangle) and the important elements of such competence, such as how knowledge is organized and how it can be made manifest. Starting with a model of learning is one of the main features that distinguishes an evidence-based approach to assessment design from more traditional approaches based on curriculum frameworks or content standards. The learning model suggests the most important aspects of student achievement about which one would want to draw inferences and provides clues about the types of assessment tasks that will elicit evidence to support those inferences (see also Pellegrino, Baxter, & Glaser, 1999; Pellegrino et al., 2001).

Consistent with these ideas, there has been growing interest in the topic of *learning progressions* (see e.g., Duschl et al., 2007; NRC, 2012; Wilson & Bertenthal, 2006). A variety of definitions of learning progressions (also called learning trajectories) now exist in the literature, with substantial differences in focus and intent (see, e.g., Alonzo & Gotwals, 2012; Corcoran et al., 2009; Daro et al., 2011; Duncan & Hmelo-Silver, 2009). Learning progressions are empirically grounded and testable hypotheses about how students' understanding of and ability to use core concepts and explanations and related disciplinary practices grow and become more sophisticated over time, with appropriate instruction (Duschl et al., 2007). These hypotheses describe the pathways that students are likely to follow as they master core concepts. The hypothesized learning trajectories are tested empirically to ensure their construct validity (Does the hypothesized sequence describe a path most students actually experience given appropriate instruction?) and ultimately to assess their consequential validity (Does instruction based on the learning progression produce better results for most students?). The reliance on empirical evidence differentiates learning trajectories from traditional topical scope and sequence specifications, typically based only on logical analysis of disciplinary knowledge and on personal experiences or preferences in teaching.

Any hypothesized learning progression has implications for assessment, because effective assessments should be aligned with an empirically grounded cognitive model. A model of a learning progression should contain at least the following elements:

1. *Target performances or learning goals*: The end points of a learning progression as defined by societal

expectations, analysis of the discipline, and/or requirements for entry into the next level of education.

2. *Progress variables*: Dimensions of understanding, application, and practice being developed by the learner and tracked over time. These may be core concepts in the discipline or practices central to literary, scientific or mathematical work.
3. *Levels of achievement*: Intermediate steps in the developmental pathway(s) traced by a learning progression. These levels, defined as points along progress variables that mark significant shifts in understanding, reflect levels of knowledge integration or common stages that characterize the development of student thinking. There may be intermediate steps that are noncanonical but are stepping stones to canonical ideas.
4. *Learning performances*: The kinds of tasks that students at a particular level of achievement would be capable of performing. They provide specifications for the development of assessments by which students would demonstrate their knowledge and understanding.
5. *Assessments*: The specific measures used to track student development along the hypothesized progression. Learning progressions include an approach to assessment, as assessments are integral to their development, validation, and use.

Research on cognition and learning has produced a rich set of descriptions of domain-specific learning and performance that can serve to guide assessment design, particularly for certain areas of reading, mathematics, and science (e.g., AAAS, 2001; Bransford et al., 2000; Duschl et al., 2007; Kilpatrick, Swafford & Findell, 2001; Snow, Burns, & Griffin, 1998; Wilson & Bertenthal, 2006). That said, there is much left to do in mapping out learning progressions for multiple areas of the curriculum in ways that can effectively guide the design of instruction and assessment. Nevertheless, a good deal is known about student cognition and learning that can be used right now to guide design of systems of assessments, especially those that attempt to cover the progress of learning within and across grades. The use of learning progression information to guide the design of an assessment serves to structure the claims made about tasks, including their appropriateness for use in an instructional context, thereby contributing information pertinent to arguments about the cognitive, instructional, and inferential aspects of the assessment's validity. The article in this issue by Bennett et al. (this issue) provides an excellent example of the application of the learning progressions framework as part of the development and validation of the CBAL assessment program (Cognitively Based Assessment of, for, and as Learning).

Assessment Development: Construct-Centered Design

The design of an actual assessment is a challenging endeavor that must be guided by theory and research about cognition in context, as well as practical prescriptions regarding the processes that lead to productive and potentially valid assessments for particular contexts of use. Design is always a complex process that applies theory and research to achieve near-optimal solutions under multiple constraints, some of which are outside the realm of science. Assessment design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a program), the context in which it will be used (e.g., classroom, district or international-comparative), and practical constraints (e.g., resources and time). The tendency in assessment design has been to work from a somewhat "loose" description of what it is that students are supposed to know and be able to do (e.g., standards or a curriculum framework) to the development of tasks or problems for them to answer. For the fractions example discussed earlier, that description might be a general statement that the assessment needs to assess students' ability to compare fractions of certain types and some constraints as to the types of tasks to use (e.g., multiple choice, short constructed response) and how many of each type, rather than the forms of evidence to be derived from any individual tasks or the overall set of problems. It is then left up to the item writer to generate some candidate items. Given the complexities of the assessment design process, it is unlikely that such a diffuse process can lead to generation of a quality assessment without a great deal of artistry, luck, and trial and error. As a consequence, many assessments are insufficient on a number of dimensions including representation of the construct, content to be covered, and uncertainty about the scope of the inferences that can be drawn from student performance on the various tasks. Assessment design under multiple constraints, as previously described, is unlikely to succeed without explicit attention to the various aspects of cognition, observation, and interpretation within an explicit domain-based context of use and interpretive purpose.

The evidentiary reasoning logic embedded in the assessment triangle is exemplified by the work of two groups of researchers who have generated frameworks for developing assessments: (a) the ECD approach developed by Mislevy and colleagues (see, e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006) and (b) the construct-modeling approach developed by Wilson and his colleagues (see, e.g., Wilson, 2005; Wilson & Draney, 2004; Wilson & Sloane, 2000). They both use a construct-centered approach to task development, and both closely follow the assessment triangle's emphasis on the logic of evidentiary reasoning.

Traditional approaches to assessment design tend to focus primarily on surface features of tasks, such as how they are presented to students or the format in which students are asked to respond. In a construct-centered approach, the selection and development of assessment tasks, as well as the scoring rubrics and criteria, and the modes and style of reporting, are guided by the construct to be assessed and the best ways of eliciting evidence about a student's proficiency with that construct.

In construct-centered approaches, the process of assessment design and development is characterized by the following developmental steps, which are common to both ECD and construct modeling:

- Analyzing the cognitive domain that is the target of an assessment.
- Specifying the constructs to be assessed in language detailed enough to guide task design.
- Identifying the inferences that the assessment should support.
- Laying out the type of evidence needed to support those inferences.
- Designing tasks to collect that evidence and modeling how the evidence can be assembled and used to reach valid conclusions.
- Iterating through the aforementioned stages to refine the elements, especially as new evidence becomes available (Pellegrino et al., 2001).

Figure 2 provides an example of the design logic in the ECD process (e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006). The figure captures three essential and interacting components of the overall process. As shown in the figure, the process starts by defining as precisely as possible the claims that one wants to be able to make about student knowledge and the ways in which students are supposed to know and understand some particular aspect of a content domain. Examples might include aspects of algebraic thinking, ratio and proportion, force and motion, heat and temperature, and so on. The most critical aspect of defining the claims one wants to make for purposes of assessment is to be as precise as possible about the elements that matter and express these in the form of verbs of cognition, for example, *compare, describe, analyze, compute, elaborate, explain, predict, justify*, and so on. Claims that use the verbs *know* or *understand* are too vague and arbitrary to guide the assessment design process in a consistent manner, because they are open to multiple forms of interpretation.

Claims about students need to be linked to the forms of evidence that would provide support for those claims—the warrants in support of each claim. The evidence statements associated with given sets of claims capture the features of work products or performances that would give substance to the claims. This includes which features need to be present and how they are weighted in any evidentiary scheme, that is, what matters most and what matters least or not at all. For example, if the evidence in support of a claim about a student's knowledge of the laws of motion is that the student can analyze a physical situation in terms of the forces acting on all the bodies, then the evidence might be drawing a free body diagram with all the forces labeled including their magnitudes and directions.

The precision that comes from elaborating the claims and evidence statements associated with the targeted aspects of a domain of knowledge and skill pays off when one turns to the design of tasks or situations that can provide the requisite evidence and related scoring rubrics. In

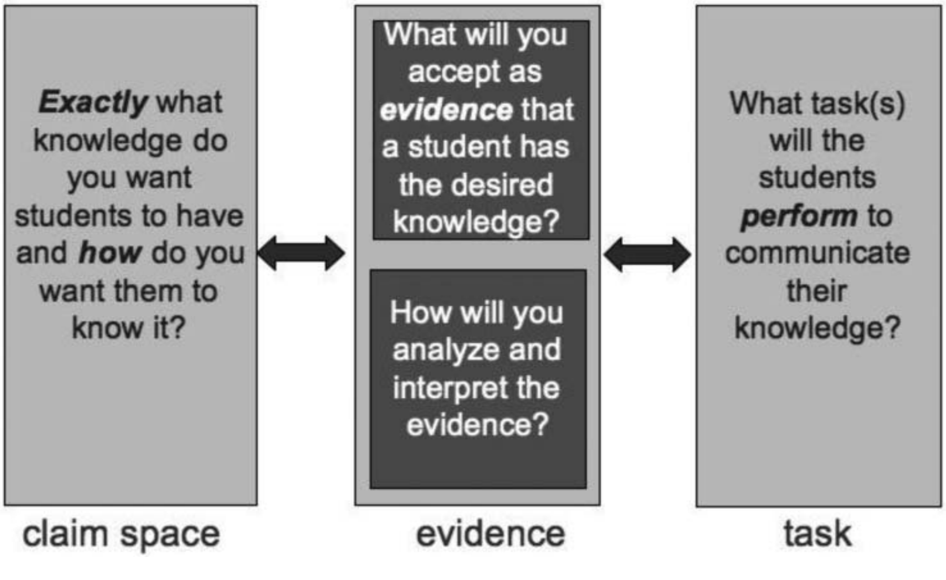

Exactly what knowledge do you want students to have and **how** do you want them to know it?
What will you accept as **evidence** that a student has the desired knowledge?
How will you analyze and interpret the evidence?
What task(s) will the students **perform** to communicate their knowledge?
claim space
evidence
task

FIGURE 2. Simplified representation of three critical components of the evidence centered design process and their reciprocal relationships.

essence, tasks are not designed or selected until it is clear what forms of evidence are needed to support the range of claims associated with a given assessment situation. The tasks need to provide all the necessary evidence, and they should allow students to "show what they know" in a way that is as unambiguous as possible with respect to what the task performance implies about student knowledge and skill, that is, the inferences about student cognition that are permissible and sustainable from a given set of assessment tasks or items. When the design logic of an assessment is instantiated in this way, elements of an argument about an assessment's cognitive, instructional, and inferential validity are instantiated in the details of the design model.

MULTIPLE FORMS OF EVIDENCE

As previously argued, assessment of student knowledge should be construed as a process of reasoning from evidence where the nature of that evidence is coupled to theories, models, and data on the nature of competence in specific areas of the curriculum and its development through processes of instruction and learning. The design of such assessments should be guided by several factors including research and theory about the nature of knowledge in the domain and the intended interpretive use of the results. A principled design process, such as ECD, can help translate theory and research about cognition and learning into an operational assessment that yields evidence aligned with the assessment's intended purpose and interpretive use. But theory, research, and principled design, combined with intuition and artistry, are not sufficient to a guarantee that an assessment will be valid relative to its particular interpretive use. In other words, no matter how well conceptualized an assessment may be relative to the cognitive constructs of interest and how principled the design process is relative to the intended interpretive use, validity of an assessment depends on retrospective evidence gathered about the completed artifact and its performance to complement the prospective argument and evidence related to its design. This is not to discount the importance of information derived from a conceptually driven and principled design processes as part of an assessment's overall validity argument but to say that further empirical evidence is needed to support claims about components of the validity of any given assessment.

Table 1 illustrates multiple possible sources of data and how they might contribute information and evidence relative to each of the three validity components. We have intentionally included a range of data sources to suggest the potential breadth and depth of validity analyses that might be conducted with assessments designed to function in proximity to classroom teaching and learning.

Each of the boxes in Table 1 provides a brief description of how that particular type of data *might* provide evidence with respect to each aspect of validity. However, whether data coming from expert analyses, or any of the other sources, do in fact provide evidence depends on the design of the protocols used for data collection and the methods of data analysis. Data collection activities need to be carefully structured so they elicit evidence relevant to the validity components intended by the assessment designer and/or researcher and so they provide forms of data that can be rigorously analyzed using both qualitative and quantitative methods.

Consider the case of expert analyses of a given set of assessment materials that are designed to help a teacher gauge her students' mastery of a mathematics topic such the addition of fractions. A group of experts, which might include mathematics educators, curriculum designers, practicing middle school teachers, and learning scientists, could be asked to judge the validity of an assessment that is part of an instructional unit in a mathematics curriculum. The assessment consists of a set of separate problems that vary the type of question asked, as well as the type of response demanded of the students from computing a value to explaining why one fraction is larger or smaller than another fraction. The experts could be asked to make several judgments including (a) the types of knowledge demanded of the student—declarative, procedural, strategic, and/or schematic—and the depth or level of sophistication at which the knowledge is probed; (b) the appropriateness of the activities given the design of the curriculum up to that point in time, including students' opportunities to develop the various forms of knowledge and understanding demanded by the tasks; (c) the information provided to the teacher as to the conditions of administration, scoring, and interpretation of performance on the tasks, as well as expectations as to the range or variability in student performance—for example, students should be able to correctly answer all the questions versus a more differentiated and diagnostic expectation regarding which items should be easy for all students and those that should be more challenging and why; and (d) the likelihood that the sets of tasks will provide evidence of mastery of relevant knowledge and skills as well as evidence of specific areas of student weakness and/or misunderstandings. These judgments would provide evidence related to one or more of the three components of validity described earlier, including the strength of the evidence for one versus another component.

We note that such judgments must be made in the context of the intended interpretive use of the assessment at that point in the curriculum and as part of the teacher's ongoing instructional practice. Thus, one of the major challenges associated with examining the validity of assessments intended to function close to instruction is that the judgments are tightly coupled to contextual factors related

TABLE 1
Relation of Data Collection Activities to Validity Components

Data Collection Activity	*Cognitive Validity*	*Instructional Validity*	*Inferential Validity*
Expert analyses	How well the design incorporates cognitively critical forms of knowledge and understanding; ethnic and cultural sensitivity review.	How well the design supports and aligns with instructional needs and uses and promotes teacher understanding.	How well the intended constructs are incorporated within the design; appropriateness of scoring rubrics and inferential models.
Student cognitive protocol studies	How well does student engagement with assessment activities correspond to design intent; how well do scoring and assessment outcomes reveal actual student thinking and proficiencies; issues regarding linguistic and cultural diversity?	How well do assessment outcomes including test and item scores interact with instructional goals and with other instructional indicators and benchmarks; how well do assessment outcomes support teacher decisions and actions?	How well does actual student engagement with assessment activities support analytic models, including model parameter interpretations, covariance analyses, and relationships to other variables?
Teacher surveys teacher logs teacher interviews	How well do teacher interpretations of student outcomes align with design intent of assessment activities; what is revealed about student understandings; what responses would be expected?	How well do teachers understand and use the assessments and assessment outcomes; how well are differential decisions and actions supported; what was teachers' actual use?	Teacher knowledge about and use of score reliabilities, item difficulties, expected student responses and variations.
Classroom observations	How sensitive are assessments to opportunities to learn relative to assessment activities; how does actual instruction support measured proficiencies?	How faithfully do teachers use assessments and what use do they make of assessment outcomes?	How sensitive is assessment performance and statistical and psychometric modeling to variability in instruction and conditions of assessment administration, and classroom uses of assessment outcomes.
Studies of item and test performance	How well do item and test performance support underlying cognitive processing demands; how well do assessment outcomes at test and item levels reflect underlying cognition?	How well do assessment outcomes support instructional needs including formative uses, summative monitoring of progress, and connections to external assessments.	How well do model-based analyses support the intended purpose and use of assessments; including scale score and diagnostic profile reliability, model-data fit, dimensionality; differential functioning for linguistic and ethnic groups; predictive validity; alignment with other tests

to a student's curricular and instructional experiences as well as teacher practice and the intended interpretive use. In fact, evaluation of the instructional validity of an assessment is very much a function of these contextual factors and the degree of coupling between and among the processes of learning, instruction, and assessment at the classroom level. In many respects it is the close coupling among these processes and the existing curriculum that differentiates the intended interpretive uses of classroom assessment from more distal large-scale assessments used by states for monitoring and accountability purposes. The latter are deliberately designed to be "neutral" with respect to curriculum and instruction.

The use of expert judgment should not be construed as some arbitrary and subjective process. Rather, expert judgment is well established in the overall literature on assessment validation, especially with respect to issues of content alignment and construct representation for many large-scale standardized assessments. It is typically done under conditions that involve protocols specifying in detail the judgment criteria to be applied and by multiple persons who have been trained to use the protocol's criteria. Their degree of agreement and consistency in doing so is systematically evaluated to determine the quality of the resultant evidence.

Consider as a second example from Table 1 the process of collecting validity evidence about an assessment by conducting student interviews or think-aloud protocols, in which a student is presented a set of assessment tasks and asked to talk about his or her thinking and reasoning while solving each task and the student is simultaneously observed by a researcher. Perhaps the student's eye movements are also being tracked, and all actions, gestures, and verbalizations during solution are being recorded for subsequent analysis. What a student says and does under such circumstances could provide evidence not only about the cognitive validity of the tasks relative to the intended cognitive constructs (and unintended ones like memory load or comprehension problems) but also about the instructional validity of the assessment in terms of the relative familiarity of the form of the assessment task given the curriculum materials and instructional practices in use in his or her classroom. Subtleties of student performance could also provide evidence related to inferential properties of the assessment such as scoring rules and procedures. As in the case of data derived from expert judgments, data derived from verbal protocols, eye movements, and/or gestures need to be structured and analyzed using rigorous procedures regarding the scope of the intended inferences and conclusions.

As a third example from Table 1, we can consider the case of collecting small-scale or large-scale samples of student performance on a given set of assessment tasks. Analyses of patterns of student performance could provide evidence relative to claims about the cognitive constructs represented in the assessment, for example, items that purportedly measure the same versus different constructs show evidence of internal structure in line with those claims. Patterns of student performance on individual items and sets of items would provide evidence of the quality of measurement of the tasks and the instrument relative to its intended score interpretation and use, for example, whether the items have appropriate and expected levels of difficulty and discriminability and whether the overall score or specific subscores are reliable. A variety of tools and methods from measurement theory and psychometrics are applicable with respect to the structuring, analysis, and interpretation of these forms of data.

APPLICATION OF THE CONCEPTUAL AND EVIDENTIARY FRAMEWORK: EVERYDAY MATHEMATICS

We have used the concepts, principles, and frameworks discussed in the three preceding sections to prospectively and retrospectively consider aspects of the validity of various examples of instructionally relevant assessments. The examples include (a) concept inventories that have increasingly come to be used in science, technology, engineering, and mathematics education settings spanning middle school through university instruction (see, e.g., the discussion by Jorion et al., 2015); (b) FACETS-based diagnostic assessments for middle school and high school physical science instruction (see, e.g., Minstrell, 2011, and the discussion by DiBello et al., in press); and (c) assessment materials embedded within K-8 mathematics and science curricula (see, e.g., the discussion by Pellegrino & Goldman, 2007). In what follows we present aspects of the latter work. In particular, we focus the discussion on analysis of the validity of assessments embedded within EDM, a popular K-6 mathematics curriculum (Bell & Bell, 2007).

Before considering our analysis of EDM, it is important to note that discussions of the design and validity of the assessments included in many popular curriculum materials are rare, and evidence is largely restricted to small-scale studies conducted by the developers of those curricula. The work presented here provides a template for an independent review of curriculum embedded assessments that goes beyond face or content validity. It illustrates the value of delving deeply into all three components of the validity of such assessments as well as some of the complexity of doing so. In the material that follows, we apply the principles of the conceptual and evidence framework presented in the prior sections to examine validity with respect to the particular uses intended for some of the different forms of assessment included in the EDM curriculum by its publishers.

Brief Overview of Curriculum, Instruction, and Assessment in EDM

In conducting validity analyses of the assessments embedded within any specific curriculum it is important to understand the structure of the curriculum and examine how the curriculum positions assessment. The EDM curriculum is designed for a relatively rapid instructional pace. At each grade level the EDM curriculum is broken up into a number of instructional units (10–12 for an entire school year). Each instructional unit consists of a number of daily lessons, and each lesson has highly structured daily routines. For example, each lesson is separated into four parts: Getting Started, Part 1: Teaching the Lesson, Part 2: Ongoing Learning and Practice, and Part 3: Differentiation Options. Teachers are encouraged to move on from a topic, even if the students have not yet mastered the topic. In fact, students are not expected to master the topic right away. Mathematical content is repeated across the school year, as topics are revisited across the year in a variety of contexts.

> Mathematical content is taught in a repeated fashion, beginning with concrete experiences. Students using EDM are expected to master a variety of mathematics skills and concepts, but not the first time they are encountered. It is a mistake to proceed too quickly from the concrete to the abstract or to isolate concepts and skills from one another or from problem contexts. Students also need to "double back" by revisiting topics, concepts, and skills and then relating them to each other in new and different ways. (Bell et al., 2007, pp. 2–3)

Everyday Math claims to have a balanced approach to assessment, which, as defined by EDM, should (a) emphasize conceptual understanding while building a mastery of basic skills; (b) explore a broad mathematics spectrum, not just basic arithmetic; and (c) be based on how students learn and what they're interested in while preparing them for their future mathematical needs (Bell et al., 2007, p. 2). Four principles form the foundation of balanced assessment in EDM: *Purpose, Context, Evidence*, and *Content*. With

TABLE 2
Specific Assessment Activities That Support the Different Sources of Evidence and Contexts for Assessment

		Context for Assessment		
		Ongoing Assessment	*Periodic Assessment*	*External Assessment*
Source of Evidence	Observation	● Informing instruction notes ● Recognizing student achievement notes for ● Mental math and reflexes ● "Kid watching"	● Progress check oral/slate assessments	● Classroom observations by resource teachers or other outside experts
	Work Product	● Recognizing student achievement notes for ● Journal slips ● Exit slips ● Game record sheets ● Writing/reasoning prompts ● Portfolio opportunities	● Mid-year and end-of-year assessments ● Progress Check written assessments ● Student self-assessments ● Open response prompts	● Standardized tests mandated by the school district or the state

respect to Purpose, assessment serves two main functions within the curriculum: formative and summative. Assessment occurs in three primary Contexts: ongoing, periodic, and external. EDM's authors suggest that the emphasis on each context should vary from greatest, to moderate, to least, respectively. Evidence about student knowledge states and reasoning capability can be indirectly gathered by "online" observation or by the student's work product. Each of the aforementioned assessment contexts can yield evidence through online observation or work product. Table 2 shows specific activities in the EDM curriculum and how they are supposed to provide evidence relevant to three contexts of assessment.

Finally, with respect to the Content being assessed, the curriculum outlines the following content strands: Number and Numeration; Operations and Computation; Data and Chance; Measurement and Reference Frames; Geometry; and Patterns, Functions, and Algebra. A set of grade-level goals is separately identified for each of these content strands. Each lesson's goals, content, and key concepts and skills are outlined at the beginning of that lesson.

As previously noted, EDM emphasizes both formative and summative functions of classroom-based assessments. Formative assessments are typically ongoing and designed to provide teachers with diagnostic information about students' current knowledge states and capabilities so that they can inform subsequent instruction. Such information is critical to differentiating instruction based on an individual student's progress. Summative assessments are periodic and give teachers indicators of student progress toward mastery of the grade-level goals. The third function—program evaluation—relates to the third context: External Assessments. External assessment refers to large-scale accountability assessments, designed to reflect state and national achievement standards. In terms of a curriculum like EDM, the external context is relevant given that policymakers and those involved in curriculum adoption decisions would like to know if there is an implied relationship between curriculum-specific assessments and the external accountability assessments.

As is evident from Table 2, EDM provides a plethora of opportunities for assessment, each of which can be evaluated with respect to aspects of cognitive, instructional, and inferential validity drawing on one or more of the five types of data sources (see Table 1). For the research we are drawing on here, we focused on in-depth analyses of a subset of assessment activities listed in Table 2 because they were associated with three content strands centrally important at the third-, fourth-, and fifth-grade levels: Number and Numeration (NN); Operations and Computation (OC); and Patterns, Functions, and Algebra (PFA). This selective focus afforded us the possibility of examining whether evidence for various forms of validity varied across years as well as for different types of assessment activities. Furthermore, while we pursued data collection for each of the five data sources, we expected that the data would be differentially informative for the three components of validity, depending on the nature of the assessment activity as well as the relative challenge of collecting data of a particular type.

Methods and data sources. At a very general level we were interested in determining whether validity evidence could be obtained for specific assessments included as part of the overall EDM curriculum design, as well as the extent of that evidence relative to each of the three components of our validity framework. Our methods of data collection and the data sources themselves were guided by the evidence structure outlined in Table 1. For example, we obtained judgments by expert reviewers of the cognitive and instructional properties of the embedded assessments relative to their mappings to curricular learning goals and the mathematics knowledge and skills. This expert group included teachers, curriculum specialists, math educators, and learning scientists. We collected three forms of behavioral data shown in Table 1: Student protocols while they performed an assessment activity, observations of classroom instruction associated with the assessment activities featured in the protocols, and teacher interviews regarding the assessments in EDM and how teachers used them (formatively, summatively, or for differentiating instruction). Student protocols consisted of observing each student when they were doing the assessment activities under instructions to say aloud what they were doing and what they were thinking. This was followed by questions that probed their conceptual understanding of the mathematics concepts tapped by the activity. In addition, students completed further tasks that drew on the same mathematical concepts under assessment task conditions different from the EDM activities. Additional tasks were selected to deepen our understanding of student knowledge and skills required by but not readily observable from the focal activities themselves. Classroom observations were intended to provide us with information about student "opportunities to learn" relative to the targeted assessment activities. The teacher interview data helped us focus on the assessment activities most commonly used in classrooms and informed selection of protocol activities, as well as the sample of assessment activities for the fifth data source: item and test performance from samples of 200 or more students. A subset of the assessment examples that were examined as part of the initial curriculum analysis were selected for large-scale data collection and statistical analysis (see Table 1). The sample of 200+ students at a given grade level was made up of all consented students in participating classrooms for that grade level across multiple years of data collection. All of these students performed each of the selected assessment examples as part of their regular classroom activities, and teachers sent student response data to the research team for scoring and analysis. For present purposes, we illustrate application of the validity analysis framework for one of the most commonly used assessment activities in EDM: unit progress checks.

Validity Analysis of a Specific Case of Assessment in EDM: Unit Progress Checks

Unit Progress Checks are one of three periodic assessment activities for which there are written student work products (see Table 2). They are administered at completion of a unit and consist of different response types (e.g., short answer, open response, and reflection). Progress Checks are designed to assess what students understand about mathematics content covered in the current and previous units. The grade-level goals tapped by the Progress Check are listed at the beginning of the lessons that make up the unit. There are two parts to the Written Progress Checks. Part A is designed to assess mastery of the specific unit goals and therefore is supposed to provide summative and diagnostic information on student progress toward the grade-level goals. Part B is intended to be used formatively and incorporates current and prior skills, as well as some content that would be learned in subsequent units. It is intended as a baseline measure for evaluating growth in the subsequent unit. Teachers could use the information gathered from that part of the assessment to plan for instruction to come.

We focus here on Part A only, to illustrate an application of the validity analysis framework. This analysis drew on two of the data sources described in Table 1: expert analyses of the items within each instrument and analyses of student performance on those same items and each instrument as a whole. As noted in Table 1, expert judgments can be used to obtain evidence related to multiple aspects of validity, and in our particular case for EDM, expert judgments were used to provide evidence relevant to instructional validity and cognitive validity. Regarding instructional validity, experts examined the representativeness of the items within a given progress check relative to the knowledge and skills targeted in the instruction for that particular unit. Regarding cognitive validity, experts examined the types and levels of knowledge tapped by the assessment tasks and the degree to which the items vary in depth of knowledge (DoK; e.g., Webb, 2007). The third form of evidence related to inferential aspects of validity is derived from analyses of large samples of student performance on the assessments. These analyses speak to measurement properties of the items and assessments as parts of inferential validity relative to the intended interpretive use.

Representativeness of instructional goals. To provide an interpretive framework for understanding the meaning of written progress check performance, the grade-level goals assessed in each written progress checks (assessed goals) were investigated in relation to the grade-level goals targeted in the unit (targeted goals). Targeted grade-level goals for a given unit are detailed in the EDM Key Concepts and Skills table under the Unit Organizer section in the *Teacher's Lesson Guide*. The Key Concepts and Skills table is intended to communicate to teachers which important mathematical ideas are covered in each lesson of the unit. A targeted grade-level goal may appear multiple times within a lesson and/or across multiple lessons in the unit.

The assessed grade-level goals in a unit's progress check were compared to the targeted grade-level goals covered in the corresponding unit. By matching assessed goals to the targeted goals covered in each unit, we are able to address the extent to which written progress checks assess each of

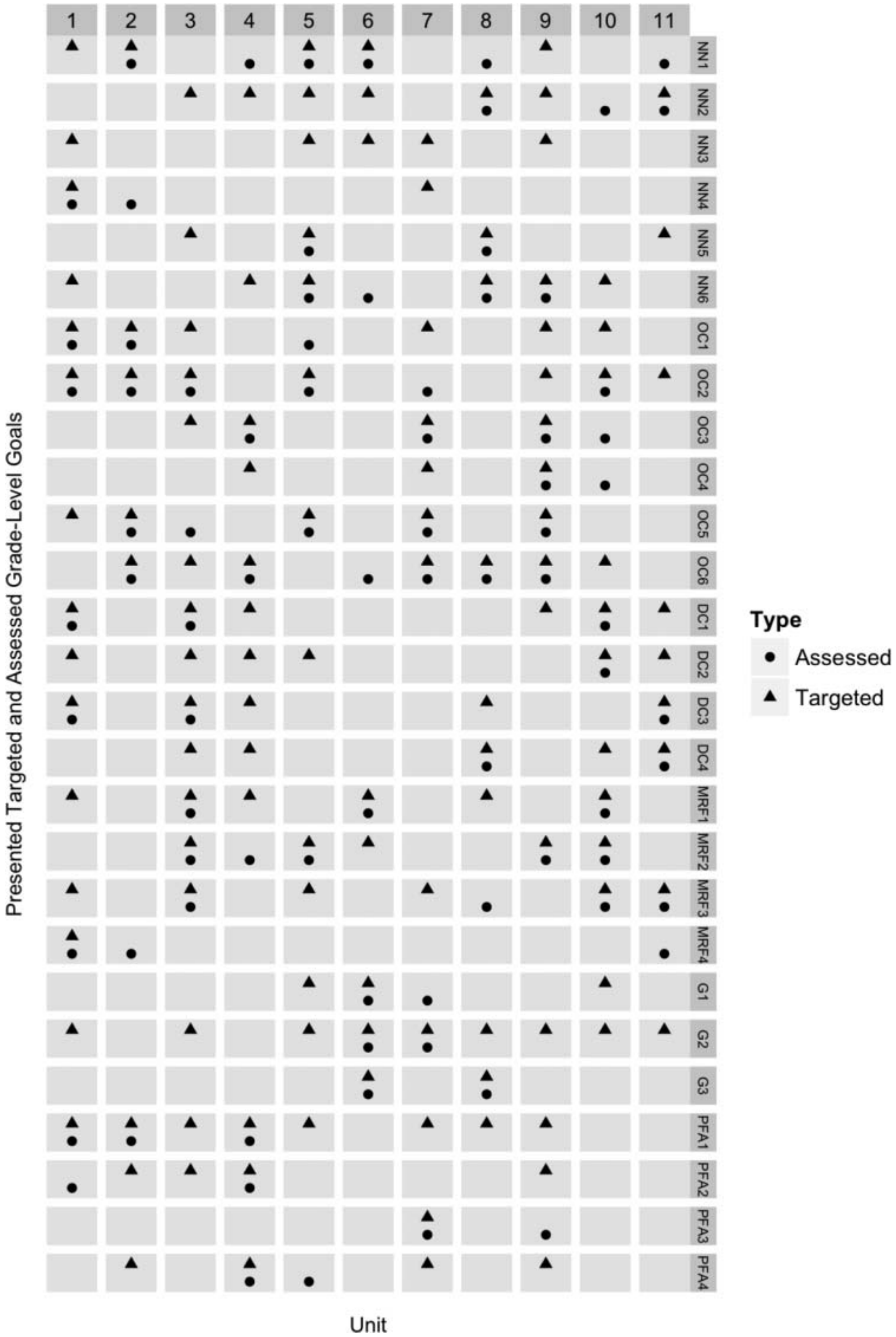


FIGURE 3 The targeted and assessed grade-level goals in each unit for Grade 3.

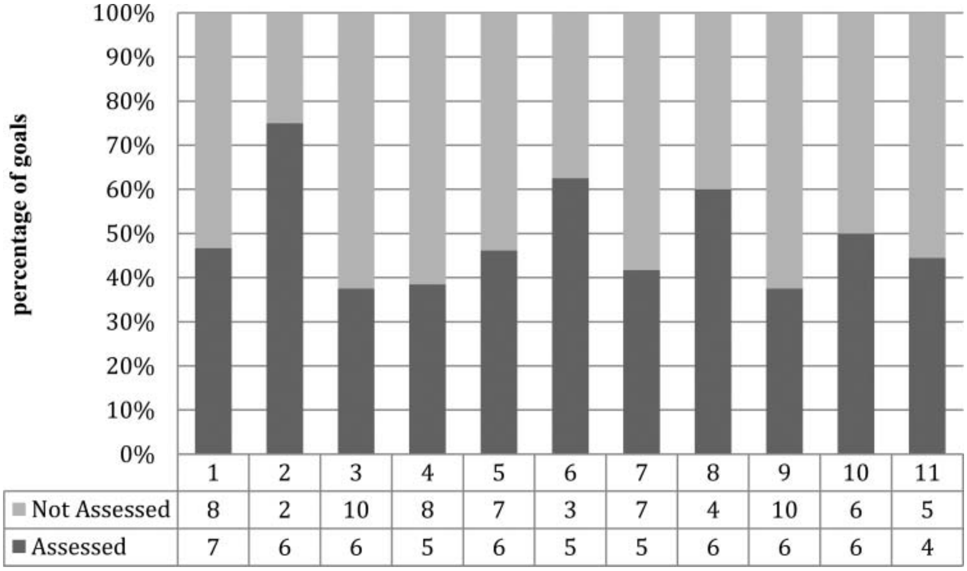
the grade-level goals covered in the corresponding unit—how many goals were assessed, with what level of concentration, and at what level of density, that is, the percentage of the targeted grade-level goals for that unit that were actually assessed. We also examined how grade-level goals assessed in written progress checks compared across units and whether it would be possible for teachers to track student progress on common goals across units. These multiple aspects of the alignment of the assessment content with the instructional goals contribute to an appraisal of the instructional validity of the Progress Checks.

Figure 3 shows where each grade-level goal for Grade 3 is targeted and where it is assessed across all units. The triangles indicate that the corresponding grade-level goal was targeted for instruction in the particular unit, and the circles indicate that the grade-level goal was actually assessed by at least one item in that unit. The grade-level goals are identified along the right column of the figure, and the units are identified along the top. Looking vertically across the grade-level goals in each unit, we see that some written progress checks assess grade-level goals that were not targeted in the corresponding units but might be targeted in previous units. For example, in Grade 3 Unit 2, NN, Goal 4 (NN4) is assessed but is not a targeted grade-level goal for this unit. NN4 was both targeted and assessed in the earlier Unit 1. In all cases, if a grade-level goal is assessed in a particular unit in which the grade-level goal was not targeted for instruction, then that grade-level goal had been targeted in an earlier unit. The single exception is the grade-level goal PFA, Goal 2 (PFA2). According to the information in Figure 3, PFA2 was first assessed in the written progress check of Unit 1 but was first targeted for instruction in Unit 2.

As can be seen by scanning down each column in Figure 3, not all of the targeted grade-level goals in a unit were assessed in the written progress check. In Grade 3 Unit 1, for instance, 15 grade-level goals were targeted. Eight of those targeted 15 were not assessed in the Unit 1 progress check written assessment: NN Goals 1, 3 and 6 (NN1, NN3, NN6); OC Goal 5 (OC5); Data and Chance 2 (DC2); Measurement and Reference Frames Goals 1 and 3 (MRF1, MRF3); and Geometry Goal 2 (G2). Figure 4 shows for each unit the total number and percentage of targeted goals for that unit broken into two parts, the targeted goals that were and were not actually measured in the unit written progress check. The bars represent the percentage of targeted goals assessed and not assessed in each unit written progress check. In Grade 3, the proportions of the targeted goals assessed ranged from a minimum of 37.5% for Units 3 and 9 to a maximum 75% for Unit 2. Three of the 11 units assessed 50% or more the targeted goals.

Returning to Figure 3, by looking horizontally across all the units in Grade 3 for each grade-level goal, it can be seen that some grade-level goals that were frequent instructional targets across the units do not appear with the same frequency in the written progress checks. Two examples are as follows: (a) Grade 3, NN3 was targeted in five units but not assessed in any of the unit written progress checks, and (b) G2 was targeted for instruction in nine of the 11 units, and only assessed in Units 6 and 7. In addition, there is limited consistency in the appearance of grade-level goals across the unit progress check assessments. Analyses of the match between targeted and assessed goals in Grades 4 and 5 showed similar patterns of co-occurrence.

The evidence from these and other analyses of the mappings of the assessment items to the grade-level goals of the units lead to three important conclusions about these written assessments. First, the degree of emphasis of assessed grade-level goals typically mismatches the emphasis of corresponding targeted instruction in the associated unit. Perhaps because of pragmatic testing constraints such as maximal acceptable testing time, the written progress checks typically do not measure a high proportion of targeted goals for a given unit, and when a particular unit's written Progress Check measures a higher proportion of that unit's targeted goals, it does so by trading off of measurement of more goals with lower adequacy of measurement (more goals = fewer items per goal). Second, based on these analyses we would not expect internal structural analyses of student performance on the written progress checks to be successful. For example, attempts to compare interitem performance correlations relative to grade-level



	1	2	3	4	5	6	7	8	9	10	11
▒ Not Assessed	8	2	10	8	7	3	7	4	10	6	5
■ Assessed	7	6	6	5	6	5	5	6	6	6	4

FIGURE 4 Percentage of targeted goals in each unit that are measured by one or more items within the written progress check for Grade 3.

TABLE 3
Definition of and Key Activities at Each DoK Level

DoK Level	*Webb's Definition*	*Key Activities*
Level 1: Recall and Reproduction	Require students to recall a simple definition, term, fact, or procedure, as well as performing a simple algorithm.	Identify, recall, recognize, use, and measure
Level 2: Skills and Concepts	Require students to make decisions on how to set up or approach a problem or activity to produce a response.	Classifying, estimating, make observations, collecting and displaying data, comparing data, determining congruent figures, and describing nontrivial patterns
Level 3: Strategic Thinking	Require students to engage in planning, reasoning, constructing arguments, making conjectures, and/or providing evidence when producing a response.	Drawing conclusions from observations; citing evidence and developing a logical argument for concepts; and using concepts to solve problems
Level 4: Extended Thinking	Require students to engage in complex planning, reasoning, conjecturing, and arguments over an extended period.	Developing and proving conjectures; making connections between a finding and related concepts

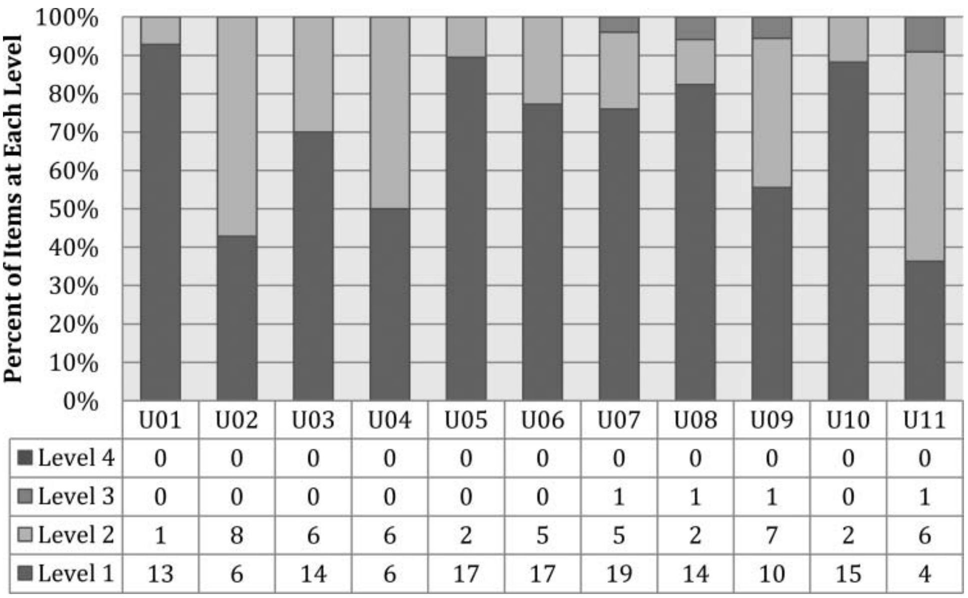
Note. DoK = depth of knowledge.

goals through either exploratory or confirmatory factor analyses were not successful. The reasons include the fact that there are relatively few items within a given written progress check that measure the particular grade-level goals. This limits the possibilities for written progress checks to provide strong diagnostic information beyond what can be determined by looking at student performance on single items. Third, the curriculum does not provide a coordinated design within and across the written progress checks that would allow teachers to track longitudinal progress on particular grade-level goals across multiple units.

Depth of knowledge. As a further part of our validity investigations, we examined the extent to which the written progress checks assess deeper learning in each unit in addition to assessing the unit's targeted goals.[1] Whether or not the progress checks imposed varying levels of cognitive demand on students and tapped multiple forms of knowledge was relevant not only to claims about their cognitive validity but also to aspects of their instructional and inferential validity. We mapped assessment items in progress checks to Webb's (2007) four-level Depth of Knowledge scale. According to Webb (1997, 2007), the DoK levels are designed to portray item content complexity that is related to cognitive processing. The four levels are determined by number of ideas integrated, depth of reasoning required, knowledge transferred to new situations, multiple forms of representation employed, and mental effort sustained. The DoK levels focus on the cognitive demand related to item content, in contrast to psychological aspects of items. That is, the DoK levels are based on a rational content analysis of what cognitive processing an item requires, rather than on an empirical analysis of how students actually approach the problem. In fact, students' strategies may be more or less complex than what an item requires. In addition, the DoK level of an item is not necessarily associated with the difficulty of an item. For example, an item at the least complex recall level could be very difficult for students. The DoK levels, definitions, and key activities are listed in Table 3 (Webb, 2007). It is worth noting that activities such as describing, explaining, and interpreting are not associated with a particular DoK level. Instead, they could fall in any DoK level depending on item format and requirements about what is to be described, explained, or interpreted.

To determine DoK levels, we used the item descriptions provided in the Everyday Math Curriculum in addition to examining the items directly. The item descriptions identify for teachers and students what the items ask a student to do, such as "fill in a table of equivalent fractions, decimals, and percents," which is similar to the description of each level's key activity in the DoK framework. In addition, similarities in the item descriptions of similar items provided an

[1] There are a variety of frameworks and approaches that can be used to evaluate the cognitive underpinnings of the tasks found on an assessment. One such approach is to have experts evaluate questions with respect to the type of knowledge required to respond to the item—*Declarative, Procedural, Schematic and Strategic* (e.g., Li, Shavelson, & White, 2002; Ruiz-Primo, 2002; Ruiz-Primo et al., 2002; Shavelson & Ruiz-Primo, 1999). Another is to use a framework such as Bloom's taxonomy, including more recent versions of that taxonomy (e.g., Anderson & Krathwohl, 2001). A third and related framework analyzes the level or depth of knowledge required by assessment tasks (Hess, 2010; Webb, 1997, 2007). Although each of these frameworks and approaches has limitations, individually and collectively they can provide useful information about the cognitive underpinnings of a given assessment and the level of challenge demanded by individual items and the collective set.



	U01	U02	U03	U04	U05	U06	U07	U08	U09	U10	U11
■ Level 4	0	0	0	0	0	0	0	0	0	0	0
■ Level 3	0	0	0	0	0	0	1	1	1	0	1
□ Level 2	1	8	6	6	2	5	5	2	7	2	6
■ Level 1	13	6	14	6	17	17	19	14	10	15	4

FIGURE 5 Distribution of Depth of Knowledge (DoK) levels across items within the progress check written assessment (PCWA) for Units 1 through 11 in Grade 3.

opportunity for keeping track of the consistency of the DoK coding. When using the DoK framework for items with subparts, each subpart was coded, and the highest DoK level across the subparts of an item was taken as the item's DoK level.

As shown in Figure 5 the vast majority of items in the unit progress checks in Grade 3 are DoK Level 1. Only two unit progress checks have 50% or more items at DoK Level 2 or higher. Those two are Units 2 and 4. Most unit progress checks have 40% or fewer items at DoK Level 2 or higher. Only four of the 11 units have any items at DoK Level 3. These four units—7, 8, 9, and 11—with some (but less than 10%) DoK 3 items are closer to the end of the school year.

The content mapping and DoK analyses of the unit written progress checks suggest the following expectations for statistical analyses of the student performance data. First, each progress check consists of sets of items that are relatively heterogeneous with regard to grade-level goals. Second, given the DoK findings and the generally small proportion of targeted skills within a unit that are actually assessed on the progress check, we would expect that students typically should demonstrate high success on most of the items. In the next section we examine the results from statistical and psychometric analyses of student performance on the Grade 3 written progress checks in pursuit of evidence about inferential validity, and in light of the evidence just described.

Analyses of student performance. Student performance on progress check written assessments was collected for all 11 units in EDM Grade 3. Table 4 provides a variety of descriptive information about each of the progress check assessments, including the number of tasks per assessment, the number of students whose data were collected and analyzed for each assessment, and three key aspects of student performance on each assessment—mean, standard deviation, and Cronbach's alpha reliability of the overall score. As indicated in the table, the number of students who provided data varied from unit to unit, and the samples only partially overlapped across units. Consequently, these data should be interpreted as reasonably representative of student performance on a given unit's assessment. However, comparison of performance across units must acknowledge the varying sample size and student composition from unit to unit.

As shown in Table 4, the progress checks varied in their composition from a low of seven to a high of 17 tasks. Some of the tasks involved multiple subparts or blanks, and for purposes of scoring these parts were scored separately and collapsed to produce a single task score. In general, student performance on the progress checks was quite good,

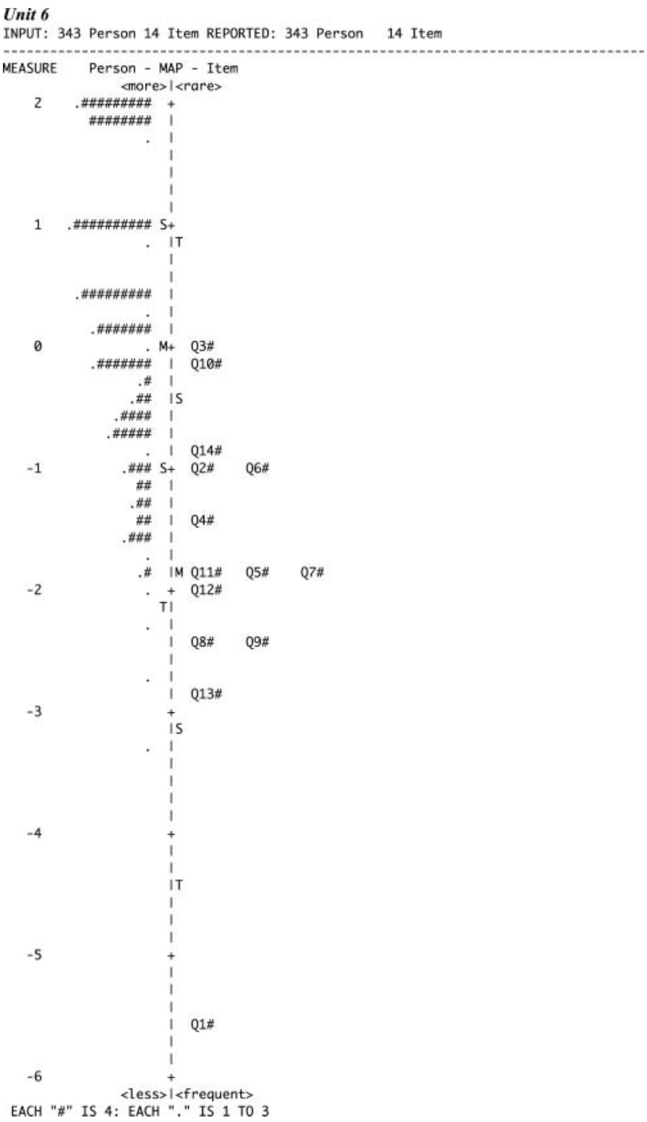
TABLE 4
Information on Grade 3 Unit Progress Checks

Unit	*No. of Tasks*	*No. of Students*	*M Performance*	*SD*	*Cronbach's α*
1	10	423	.74	.13	.66
2	10	450	.77	.17	.73
3	10	326	.66	.22	.64
4	8	378	.85	.16	.72
5	12	360	.81	.16	.68
6	14	343	.79	.17	.73
7	17	352	.87	.10	.69
8	10	360	.70	.17	.68
9	15	335	.65	.18	.73
10	14	215	.67	.18	.72
11	7	150	.56	.23	.61

with several progress checks showing average performance well above 0.7. As implied by the standard deviations of student performance on each unit progress check, the distribution of student scores was generally negatively skewed.

At the level of individual tasks, many of the item *p* values were greater than 0.8, that is, average task scores across all students were greater than 0.8. Such results are consistent with the DoK analysis reported earlier suggesting that the degree of cognitive challenge was low for the majority of the items found on each progress check. Also shown in Table 4 are the overall reliabilities of the progress checks. The Cronbach's alpha scores were in the low to moderate range for assessments of this length.

Given the sample sizes we also conducted analyses of performance using item response theory. Rasch analyses were conducted for each progress check, and these generally revealed a weak alignment between estimates of student proficiency and item difficulty levels. Such poor alignment can be seen in the Wright map shown in Figure 6 for the Unit 6 progress check. The Wright map places student proficiencies (shown on the left) and item difficulty parameters (shown on the right) on the same scale. The distribution of estimated student proficiencies is shifted upward on the scale, whereas the distribution of item difficulties is shifted downward on that same scale. In general, the items concentrate at a lower level of challenge than the estimated achievement of the majority of the students. This produces a peaked test information function centered at a low level on the logit scale (see Figure 6, where the peak is near −2.0 logits). The IRT results also suggest why the Cronbach's alpha scores are modest—many items have high average scores across students and therefore demonstrate low discriminability relative to variation in the total score.



Unit 6

```
INPUT: 343 Person 14 Item REPORTED: 343 Person   14 Item
---------------------------------------------------------------------------------
MEASURE    Person - MAP - Item
               <more>|<rare>
    2    .#########  +
           ########  |
                  .  |
                     |
                     |
                     |
                     |
    1   .########## S+
                  .  |T
                     |
                     |
         .#########  |
                  .  |
           .#######  |
    0             . M+  Q3#
           .#######  |  Q10#
                 .#  |
                .##  |S
              .####  |
             .#####  |
                  .  |  Q14#
   -1          .### S+  Q2#    Q6#
                 ##  |
                .##  |
                 ##  |  Q4#
               .###  |
                  .  |
                 .#  |M Q11#   Q5#    Q7#
   -2             .  +  Q12#
                    T|
                  .  |
                     |  Q8#    Q9#
                     |
                  .  |
                     |  Q13#
   -3                +
                     |S
                  .  |
                     |
                     |
                     |
                     |
   -4                +
                     |
                     |
                     |T
                     |
                     |
                     |
   -5                +
                     |
                     |
                     |
                     |  Q1#
                     |
                     |
   -6                +
               <less>|<frequent>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

FIGURE 6 Wright Map for performance on Grade 3 Unit 6 progress check.

Summary interpretation of the evidence. Based on the content mapping, DoK analysis, and student performance analyses, the unit progress check written assessments appear to be designed to function more like a criterion-referenced measure at a low level of cognitive challenge. This is consistent with a purpose of indicating whether students pass a minimal threshold rather than differentiating students with respect to overall unit proficiency. In fact, the maximal information for purposes of discriminating levels of student achievement appears to be at relatively low levels of student performance. This is illustrated in Figure 7 for Unit 6, where the test information peaks around scale value near −2.0, which is a relatively low proficiency level. Given the depth of knowledge findings and the typically small proportion of targeted skills within a unit actually assessed on the progress check, the criterion level measured by the assessment is something like minimal proficiency for students to be considered as having made adequate progress within each unit.

What, then, do these results imply relative to issues of cognitive, instructional, and inferential validity for an assessment with the purpose of providing ongoing information about student progress toward attaining key curricular and instructional goals in mathematics at a given grade level? If the interpretive use of a progress check is to ascertain whether individual students have mastered relevant cognitive content that had been the focus of instruction in the preceding unit (as well as some content from prior units), then one would have to conclude that the instrument provides weak evidence of level of mastery beyond very basic forms of mathematical knowledge and skill. The level of cognitive challenge of the tasks is not great, the coverage of goals targeted in the unit is low, and the structure of the instrument does not support generation of identifiable and reliable subscores for specific targeted mathematical goals.

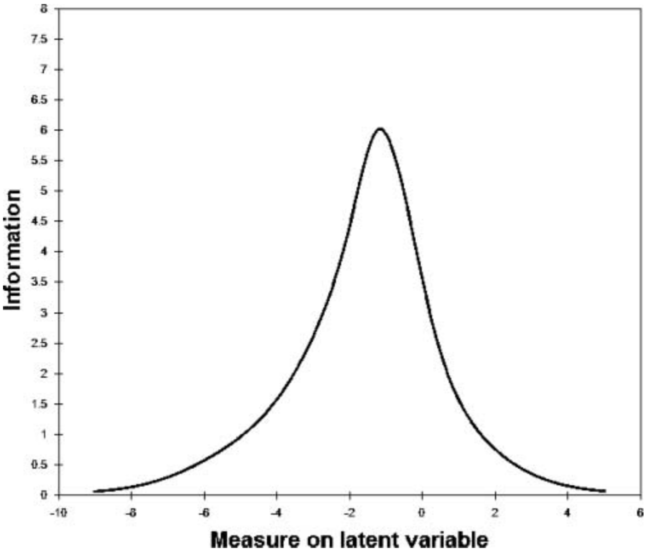


FIGURE 7 Test information function for Grade 3 Unit 6 progress check.

The cognitive diagnostic capacity is limited by the structural design features such as high variability across units in what aspects of the mathematical goals are represented and their absolute and relative frequency and density.

With respect to inferential validity, the progress checks appear to be limited in their measurement properties in several ways. In addition to limited diagnostic capacity relative to specific areas of mathematics knowledge and skill, they do not do a precise job of differentiating among students in terms of overall levels of proficiency, except at the lower ends of performance. For purposes of making a summative judgment and possibly using the progress check total score for grading, the instruments have limited reliability with respect to the total score.

With respect to instructional validity, there are several facets of the analysis that question the strength of the evidence relative to this validity component. On one hand, the assessments do cover mathematical content that was the focus of the given instructional unit, but the nature of that coverage is inconsistent from unit to unit with respect to which of the targeted instructional goals are represented and how the items are distributed across those goals that are actually assessed. With respect to diagnostic use of performance for purposes of instructional planning, the evidence suggests that other than identifying students who are struggling with the content overall, Part A of the instruments has limited instructional planning value.

All of the foregoing judgments about aspects of validity of the unit progress checks must ultimately be viewed in context: What does the curriculum intend with respect to teacher use of the results of these instruments, and how well do the instruments function with respect to those intended interpretive uses of performance on the instruments? Thus, we are not making an absolute judgment about the validity of these assessments within the EDM curriculum but are offering evidence relative to various aspects of validity with respect to a range of plausible interpretive uses. The evidence we have obtained relative to aspects of cognitive, instructional, and inferential aspects of validity for these assessments also suggests ways in which the instruments might be redesigned to enhance one or more of the possible interpretive uses. For example, because adequate coverage of all the instructional goals of a unit might lead to too long of an assessment for administration in a typical class period, the assessment might be focused instead on a key subset of those goals with appropriate variation in the type of knowledge and DoK tapped by tasks related to those goals to yield student-level diagnostic information on the progress of students in attaining important mathematical knowledge and skills. The mathematical foci of the progress checks might also be varied in a systematic way across the unit progress checks consistent with one or more progress variables of interest.

Any possible redesign, however, requires considerations of multiple constraints impacting the design and use of these assessments. Such constraints include time allocated to test administration, ease of scoring, and interpretation of performance on these assessments, as well as how they are intended to function relative to the various other assessments and procedures that EDM includes as part of its overall curricular, instructional, and assessment design. In particular it is important to remember that the question of how these assessments might be redesigned in order to improve aspects of their validity relative to their intended uses cannot be separated from the necessity to look at design of the instructional features of the curriculum and how they are related to the assessments. By its nature, curriculum-embedded assessment often blurs the line between assessment and instruction. A deep investigation of validity of a curriculum's embedded assessments is a useful first step for looking at the design intents of the curriculum itself, the sequence of goals and activities, how they are operationalized for teachers and learners, and how assessment can be coordinated with instruction to improve both instructional practice and student learning.

Finally, in focusing on Part A of the written progress checks in EDM, we have provided but one possible example of the process of pursuing validity analyses of assessments embedded in a curriculum. No judgment should be made from this one example about the validity of each of the range of assessment types and tasks found within EDM, or their collective validity. In fact, our research on the validity of assessment in EDM has provided evidence that supports various aspects of the cognitive, instructional, and inferential aspects of validity of different tasks in EDM relative to their intended interpretive uses. Another example of our analyses of EDM assessments intended to monitor progress in attaining the grade-level curricular goals can be found in our exploration of the midyear and end-of-year assessments. In contrast to some of the inferential validity evidence obtained for the progress checks, these instruments appeared to have more desirable measurement properties, perhaps because they included more items and a broader sampling of the curricular goals. The range of item difficulties better aligned with the range of student performance, perhaps because the instrument was probing retention of knowledge that was less proximal to instruction than had been the case for the unit progress checks. One additional aspect of the validity of these other instruments, which may or may not have been intended by the curriculum developers, was that performance on these assessments tended to correlate reasonably well with performance on the state's large-scale achievement test for mathematics.

FURTHER CONSIDERATIONS AND FUTURE DIRECTIONS

Educational assessment is a complicated enterprise, and much of the discussion regarding assessment tends to focus on large-scale standardized tests, especially given their prominent use in the United States for purposes of accountability. Similarly, much of the literature on the validity of educational assessments has tended to focus on aspects of validity as applied to the interpretive uses of such instruments. Only within the last 15 years has there been significant effort to balance the discussion to acknowledge that assessment that functions close to classroom teaching and learning requires greater attention in terms of its importance for attaining educational goals, especially relative to the emphasis on large-scale achievement testing (see, e.g., Gordon Commission on the Future of Assessment in Education, 2013a, 2013b). Even so, relatively little attention has been given to the design and validation of such assessments.

Although assessments are currently used for many purposes in the educational system, a premise of the *Knowing What Students Know* report (Pellegrino et al., 2001) is that their effectiveness and utility must ultimately be judged by the extent to which they promote student learning. The aim of assessment should be "*to educate and improve* student performance, not merely to *audit* it" (Wiggins, 1998, p. 7). Because assessments are developed for specific purposes, the nature of their design is very much constrained by their intended use. The reciprocal relationship between function and design leads to concerns about the inappropriate and ineffective use of assessments for purposes beyond their original intent. To clarify some of these issues of assessment purpose, design, and use, it is worth considering two pervasive dichotomies in the literature that are often misunderstood and conflated.

The first dichotomy is between *internal* classroom assessments administered by teachers and *external* tests administered by districts, states, or nations. Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) showed that these two very different types of assessments are better understood as two points on a continuum that is defined by their distance from the enactment of specific instructional activities. They defined five discrete points on the continuum of assessment distance: *immediate* (e.g., observations or artifacts from the enactment of a specific activity), *close* (e.g., embedded assessments and semiformal quizzes of learning from one or more activities), *proximal* (e.g., formal classroom exams of learning from a specific curriculum), *distal* (e.g., criterion-referenced achievement tests such as required by the U.S. No Child Left Behind legislation), and *remote* (broader outcomes measured over time, including norm-referenced achievement tests and some national and international achievement measures). Different assessments should be understood as different points on this continuum if they are to be effectively aligned with each other and with curriculum and instruction.

A second pervasive dichotomy is the one between formative assessments used to advance learning and summative assessments used to provide evidence of prior learning. Often it is assumed that classroom assessment is synonymous with formative assessment and that large-scale assessment is synonymous with summative assessment. What are now widely understood as different types of assessment practices are more productively understood as different functions of assessment practice, and summative *and* formative functions can be identified for most assessment activities, regardless of the level on which they function.

Drawing from the work of Lemke (2000), it is apparent that different assessment practices can be understood as operating at different *timescales*. The timescales for the five levels just defined can be characterized as *minutes, days, weeks, months*, and *years*. Timescale is important because the different competencies that various assessments aim to measure (and, therefore, the appropriate timing for being impacted by feedback) are *timescale specific*. The cycles, or periodicity, of educational processes build from individual utterances in the moment to overall knowledge and competency represented across an individual's lifespan of educational development. What teachers and students say in class constitute verbal exchanges, these exchanges make up the lesson, a sequence of lessons make up the unit, units form a curriculum, and the curricula form an education. Each of these elements operates on different cycles or timescales: second to second, day to day, month to month, and year to year.

The level at which an assessment is intended to function, which involves varying distance in "space and time" from the enactment of instruction and learning, has implications for how and how well it can fulfill various functions of assessment, be they formative, summative, or program evaluation (see NRC, 2003). As argued elsewhere (Hickey & Pellegrino, 2005; Pellegrino & Hickey, 2006), it is also the case that the different levels and functions of assessment often have varying degrees of match with theoretical stances about the nature of knowing and learning such as the *behaviorist, cognitive*, and *situative* perspectives as discussed by Greeno, Collins, and Resnick (1996). These perspectives are not mutually exclusive, but they emphasize different aspects of knowing and learning with differing implications for what should be assessed and how the assessment process should be transacted (see, e.g., Greeno et al., 1996a; Greeno, Pearson, & Schoenfeld, 1996). Assessment practices at the immediate, proximal, and remote levels prototypically are better aligned with sociocultural, rationalist, and empiricist assumptions, respectively, whereas the close and distal levels are intermediate and align with hybrid sociocultural/rationalist and rationalist/empiricist assumptions, respectively. This assignment of perspectives to levels is based on a shift across levels in the degree of social context and mediation of the assessment activities and in the degree of remoteness or abstraction from the original processes of learning and instruction with respect to the tools, artifacts, discourse structures, and persons involved (see Hickey & Pellegrino, 2005; Pellegrino & Hickey, 2006, for a further explanation of these considerations).

Regardless of the level at which an assessment is intended to function and the framing of the assessment activities in terms of theories and models of learning and knowing, issues of validity remain paramount. We would argue that if assessments are intended to have an “educative” function and to support ongoing processes of teaching and student learning, then the three components of *cognitive, instructional*, and *inferential validity* are relevant and applicable to the conduct of a validity analysis, whatever the assessment’s intended purpose and/or level of functioning relative to ongoing processes of teaching and learning. The challenge of course is to first determine what claims are explicit or implicit about the design and use of a given assessment and then determine the forms of evidence that would be relevant to substantiating such claims relative to the intended interpretive use(s).

In this article we have illustrated how such a validity argument might be constructed for one type of instructionally supportive assessment. Despite all the press associated with external assessments in the educational system, the most frequent use of assessment is in the classroom. Teachers rely on a variety of assessment tools and resources as a part of their everyday practice. Yet much of what they use has limited validity evidence at best. Many of those materials come from publishers of curriculum materials and other instructional resources. It is time to apply clear standards of validity evidence to such materials and demand that validity analyses be conducted for such assessment materials. It is critical that educators and the public know whether the materials provided to teachers and students support or undermine effective practices of teaching and learning. In addition, as suggested by the results found here, an incorporation of results from an empirically and theoretically based validity evaluation of curriculum embedded assessments has the capacity to improve the design and development of curriculum assessment activities and the instructional materials and actions they can inform. This may be especially pertinent given significant shifts in the goals for learning such as those signaled by the Common Core Standards in Mathematics and English Language Arts (Common Core Standards in Mathematics and English Language Arts, 2010a, 2010b) and the Framework for K-12 Science Education and the Next Generation Science Standards (Achieve, 2013; NRC, 2012).

FUNDING

Development of the conceptual framework described in this article and its application to assessments intended for classroom use was supported in part by grants to the authors from the Institute of Education Sciences, U.S. Department of Education (#R305A100475, #R305A090111), and the National Science Foundation (#0732090, #0918552, #0920242). The opinions expressed are those of the authors and do not represent views of the funding agencies.

REFERENCES

Achieve. (2013). *Next generation science standards*. Retrieved from http://www.nextgenscience.org/

Alonzo, A. C., & Gotwals, A. W. (2012). *Learning progression in science: Current challenges and future directions*. Rotterdam, The Netherlands: Sense Publishers.

American Association for the Advancement of Science. (2001). *Atlas of science literacy*. 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: American Educational Research Association. (Prior version published 1999)

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom’s taxonomy of educational objectives*. Boston, MA: Allyn & Bacon.

Bell, M., & Bell, J. (2007). *Everyday mathematics: The University of Chicago school mathematics project* (3rd ed.). Chicago, IL: McGraw-Hill Wright Group.

Bell, M., Bell, J., Bretzlauf, J., Dillard, A., Flanders, J., Hartfield, R., ... Saecker, P. (2007). *Everyday mathematics teacher’s reference manual grades 1–3*. Chicago, IL: McGraw-Hill Wright Group.

Bennett, R. E., Deane, P.W., & van Rijn, P.W. (this issue). From cognitive-domain theory to assessment practice. *Educational Psychologist*, *51*(1).

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). *Assessment for learning: Putting it into practice.* Maidenhead, UK: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, *5*(1), 7–73.

Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.). (2000). *How people learn: Brain, mind, experience, and school* (expanded ed.). Washington, DC: National Academies Press.

Common Core State Standards Initiative. (2010a). *English language arts standards*. Washington, DC: National Governors Association and Council of Chief State School Officers. Available from http://www.corestandards.org/the-standards/english-language-artsstandards.pdf

Common Core State Standards Initiative. (2010b). *Mathematics standards*. Washington, DC: National Governors Association and Council of Chief State School Officers. Available from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Corcoran, T. B., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. New York, NY: Columbia University, Teachers College, Consortium for Policy Research in Education, Center on Continuous Instructional Improvement.

Daro, P., Mosher, F. A., Corcoran, T., Barrett, J., & Consortium for Policy Research in Education. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. Philadelphia, PA: Consortium for Policy Research in Education.

DiBello, L. V., Pellegrino, J. W., Gane, B. D., & Goldman, S. R. (in press). A validity analysis framework for instructionally supportive assessments. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments*. New York, NY: Routledge.

DiBello, L. V., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. V. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume on psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, The Netherlands: Elsevier B. V.

Duncan, R. G., & Hmelo-Silver, C. (2009). Learning progressions: Aligning curriculum, instruction, and assessment. *Journal for Research in Science Teaching*, *46*, 606–609.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grade K-8*. Washington, DC: The National Academies Press.

Gordon Commission on the Future of Assessment in Education. (2013a). *A public policy statement*. Princeton, NJ: Author. Available from http://www.gordoncommission.org/publications_reports.html

Gordon Commission on the Future of Assessment in Education. (2013b). *To assess, to teach, to learn: A vision for the future of assessment* (Technical report). Princeton, NJ: Author. Retrieved from http://www.gordoncommission.org/publications_reports.html

Greeno, J. G., Collins, A. M., & Resnick, L. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 15–46). New York, NY: MacMillan.

Greeno, J. G., Pearson, P. D., & Schoenfeld, A. H. (1996, August). *Implications for NAEP of research on learning and cognition*. Stanford, CA: National Academy of Education.

Haertel, E. H., & Lorié, W. A. (2004). Validating standards-based test score interpretations. *Measurement*, *2*, 61–103.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Thousand Oaks, CA: Corwin Press.

Hess, K. K. (2010). *Applying Webb's depth of knowledge levels in science*. Dover, NH: National Center for the Improvement of Educational Assessment. Retrieved from http://www.nciea.org/beta-site/publication_PDFs/DOKscience_KH11.pdf

Hickey, D., & Pellegrino, J. W. (2005). Theory, level, and function: Three dimensions for understanding transfer and student assessment. In J. P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 251–293). Greenwich, CO: Information Age.

Jorion, N., Gane, B., James, K., Schroeder, L., DiBello, L., & Pellegrino, J. (2015). An analytic framework for evaluating the validity of concept inventory claims. *Journal of Engineering Education*, *104*, 454–496.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, *38*, 319–342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement*, *21*, 31–41.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.

Kilpatrick, J., Swafford, J., & Findell, B. (Eds.). (2001). *Adding it up: Helping children learn mathematics*. Washington, DC: National Academies Press.

Lemke, J. J. (2000). Across the scale of time: Artifacts, activities, and meaning in ecosocial systems. *Mind, Culture, and Activity*, *7*, 273–290.

Li, M., Shavelson, R. J., & White, R. T. (2002). *Toward a framework for achievement assessment design: The case of science education*. Stanford, CA: School of Education, Stanford University.

Marion, S., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, Winter*, *25*, 47–57.

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

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

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley (Ed.), *Designing for science: Implications from everyday, classroom, and professional settings* (pp. 415–443). Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33*, 379–416.

Mislevy, R. J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, *25*, 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L., & Almond, R. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment*. Washington, DC: National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas* (Committee on a Conceptual Framework for New K-12 Science Education Standards, Board on Science Education). Washington, DC: National Academies Press.

Partnership for Assessment of Readiness for College and Careers. (2014). *The PARCC assessment: Item development*. Information available at http://www.parcconline.org/assessment-development

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (Vol. 24, pp. 307–353). Washington, DC: American Educational Research Association.

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

Pellegrino, J. W., & Goldman, S. R. (2007). Beyond rhetoric: Realities and complexities of integrating assessment into teaching and learning. In C. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 7–52). Mahwah, NJ: Erlbaum.

Pellegrino, J. W., & Hickey, D. (2006). Educational assessment: Towards better alignment between theory and practice. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology:*

Past, present and future trends. Sixteen essays in honour of Erik De Corte (pp. 169–189). Oxford, UK: Elsevier.

Pellegrino, J. W., Wilson, M., Koenig, J., & Beatty, A. (Eds.). (2014). *Developing assessments for the next generation science standards*. Washington, DC: National Academies Press.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393.

Smarter Balanced Assessment Consortium. (2014). *The SBAC assessment: Item writing and review*. Information available at http://www.smarterbalanced.org/smarter-balanced-assessments/item-writing-and-review/.

Shavelson, R. J. & Ruiz-Primo, M. A. (1999). On the assessment of science achievement. (English version) *Unterrichts wissenschaft*, *27*, 102–127.

Snow, C. E., Burns, M., & Griffin, M. (Eds). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academies Press.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York, NY Springer.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (National Institute for Science Education and Council of Chief State School Officers Research Monograph No. 6). Washington, DC: Council of Chief State School Officers.

Webb, N. L. (2007). Mathematics content specification in the age of assessment. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 1281–1292). Charlotte, NC: Information Age.

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

Wiliam, D. (2007). Keeping learning on track: formative assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age.

Wiliam, D. (2012). *Embedded formative assessment: Practical strategies and tools for K-12 teachers*. Bloomington, IN: Solution Tree Press.

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

Wilson, M. R., & Bertenthal, M. W. (Eds.). (2006). *Systems for state science assessments*. Washington, DC: National Academies Press.

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR Assessment System. *Yearbook of the National Society for the Study of Education*, *103*, 132–154.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*, 181–208.