3

Principled Approaches to Assessment Design, Development, and Implementation

Steve Ferrara, Emily Lai, Amy Reilly, and Paul D. Nichols

Long ago, John Bormuth referred to the process of item and test development as a "dark art," in which "construction of achievement test items [is] defined wholly in the private subjective life of the test writer" (Bormuth, 1970, pp. 2–3; also cited in Ferrara, 2006, p. 2). Much has changed – or, rather, much is in the process of changing. Some assessment programs now use principled approaches to assessment design, development, and implementation that shed light on the "dark art." Similarly, many assessment programs now use an argumentation approach to the validation of test score interpretations and uses (see Kane, 2006, 2013, 2016), though the matter of how to implement this approach in a consistent and rigorous manner is far from settled (see Borsboom & Markus, 2013; Lissitz & Samuelson, 2007).

In this chapter, we describe and develop five **foundation elements** and an **organizing element** that define principled approaches to assessment design, development, and implementation and the ongoing accumulation and synthesis of evidence to support claims and **validity arguments**. Specifically, the five foundation elements are (a) clearly defined **assessment targets**, (b) a statement of **intended score interpretations and uses**, (c) a **model of cognition, learning, or performance**, (d) aligned **measurement models** and reporting scales, and (e) **manipulation of assessment activities** to align with assessment targets. The overarching, organizing element is the ongoing accumulation of evidence to support validity arguments.

We illustrate five **principled assessment design** approaches currently in use that adapt and embed the foundation elements and discuss how the five approaches emphasize the five elements differently. The five approaches are:

1. **Evidence-centered design** (ECD),
2. **Cognitive design systems** (CDS),
3. **Assessment engineering** (AE),

The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications, First Edition.
Edited by André A. Rupp and Jacqueline P. Leighton.

4. **Berkeley Evaluation and Assessment Research** (BEAR) **Center assessment system** (BAS), and
5. **Principled design for efficacy** (PDE).

We discuss these five principled approaches with large scale educational achievement assessment purposes primarily in mind. However, our discussion is also relevant to the design, development, and implementation of interim, benchmark, and formative educational assessments and to psychological, workplace, and credentialing tests.

We argue that readers interested in making examinee cognition explicit in assessment activities and test-score interpretations should seriously consider principled approaches to assessment design, development, and implementation because the evidentiary demands for creating validity arguments for such assessments are often more robust than they are for conventional achievement tests. Specifically, assessments intended to support inferences about examinee cognition typically are designed using a theory of cognition. Such theory-based interpretations carry the extra burden of demonstrating evidence that can link assessment performance to the construct as it is defined by the theory, including all of the claims subsumed within that theory (Kane, 2006). Principled approaches help to make articulation of the claims in this complex chain explicit and accumulation of evidence to support the claims a procedural by-product.

Saying that these five approaches are "principled" should not be taken to mean that traditional and current assessment design, development, and implementation practices are "unprincipled". Rather, we mean that, relative to principled approaches, the six elements play a less explicit role in traditional and current practice and may exist in test designers' and item writers' heads rather than in **item** specifications, as Bormuth observed in 1970. Further, rationales for design decisions often are based on practical constraints (e.g., "we can afford only 90 minutes of testing time, which allows for 40 **multiple choice** items and five short **constructed-response** items") at the expense of supporting inferences about deep learning and higher order thinking that may be intended or desired. In fact, rationales for design decisions often are not well documented at all; see the discussion in Ferrara and Lai (2016, pp. 606–611).

Similarly, throughout the chapter, we refer to the processes of design, development, and implementation. By "design" we mean decisions about intended test-score interpretations and uses, numbers and types of assessment activities, testing time, delivery mode, various ways of specifying assessment activities (e.g., item specifications, item templates) to guide item writers, and related decisions. By "development" we mean activities in which assessment specifications are fulfilled by people or computer programs to produce assessment activities plus review, revision, and approval activities, including pilot and field testing and committee reviews of items, and test forms assembly.

Finally, by "implementation" we refer broadly to operational administration of assessments to examinees, including processes such as response scoring, item and test analysis, scaling, equating, score reporting, and documentation of test technical characteristics and validity arguments (see Ferrara & Lai, 2016). We also mean it quite specifically, as in the principle of fairness in testing (i.e., fair treatment during the testing process, which includes both standardization and flexibility) in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, chapter 3).

We have organized this chapter as follows. We first provide a few additional rationales for the consideration of principled assessment design approaches. Then, in a first main section, we propose and discuss five conceptual foundation elements and an organizing element that characterize each approach. In the second main section, we review the five principled approaches using these six elements. Specifically, we demonstrate that all five approaches are similarly principled but reflect principled elements in their own unique ways. In an extended discussion section, we discuss how these approaches are similar to, and different from, conventional approaches to principled assessment design. We also reflect on practical challenges and considerations for selecting a principled approach and end with a few speculations about the evolution of principled approaches in the future.

Motivations for Principled Approaches

We are not aware of specific job description components for the position of an "assessment designer" that relate closely to evidentiary reasoning practices nor of graduate programs focused specifically on the design of educational, psychological, workplace, and credentialing tests. This is quite different from other fields such as architecture, aeronautics, and software engineering where principled design is standard practice. In educational measurement, psychometricians, subject matter experts, and even policy makers commonly specify numbers and types of test items, testing time limits, mode of administration, and other design elements. Design engineers in other fields, however, constantly develop new products and processes to achieve *functional utility* (see, e.g., https://en.wikipedia.org/wiki/Design_engineer). In assessment design, the goal is likewise to achieve functional utility. In this context, functional utility requires the provision of evidence to support intended inferences about what examinees know, can do, have achieved, or their current level of development toward proficiency or expertise in an academic content area, field of endeavor, or on a psychological construct. As in all design contexts, assessment design is a process of maximizing within constraints (H. Braun, personal communication, April 4, 2014).

Principled assessment design, development, and implementation is a "logical, systematic approach to test creation" (from Zieky, 2014, p. 79, referring specifically to ECD). We use the term principled "approaches" because this broad term subsumes the conceptual and action-oriented definitions of related terms such as "system" and "framework."[1] Principled approaches in educational assessment can be characterized by their focus on validity arguments that support intended score interpretations and uses and development of empirical evidence to support those score interpretations and uses throughout the design, development, and implementation process.

Principled approaches provide concepts, procedures, and tools to guide assessment design, development, and implementation decisions. These tools are intended to align all design elements in an assessment: assessment targets, assessment activities and response scoring, measurement models, and test scores and intended interpretations and uses of test scores. They require the use of empirical evidence, where it exists, and arguments or rationales (i.e., theory, conceptual analysis, or logic) to support decisions that align these elements. Moreover, they require test developers and testing program

managers to capture evidence throughout the design, development, and implementation process to support score interpretations and uses – that is, to support validity arguments. Every step in the process, every decision, is subject to review, iteration, and re-cycling throughout the design, development, and implementation process. Re-cycling is intended to refine the alignment among all elements of the assessment process and the evidence that supports the validity argument.

One might reasonably ask why we have frameworks for principled approaches at the time of this writing, why testing programs are implementing them, and why some assessment programs are moving beyond more traditional practices more than others. As we suggested in the opening of this chapter, there exists a long standing concern that we do not know a lot about item writer thinking (e.g., Bormuth, 1970). The *National Research Council* cited another shortcoming in current assessment design and development practices:

> The central problem … is that most widely used assessments of academic achievement are based on highly restrictive beliefs about learning and competence not fully in keeping with current knowledge about human cognition and learning. Likewise, the observation and interpretation elements underlying most current assessments were created to fit prior conceptions of learning and need enhancement to support the kinds of inferences people now want to draw about student achievement. (National Research Council, 2001, pp. 2–3)

Several recent intellectual and technical developments can be cited as influences on how to respond to these shortcomings. These include recognition that learning and other social sciences have much to teach us about learning, how it develops, and how to assess it (e.g., Mislevy, 2006; Snow & Lohman, 1989), the rise of validity argumentation as the prevailing view on assessment validation (Kane, 2006, 2013, 2016), the call in *Knowing What Students Know* (National Research Council, 2001, p. 2) for due attention to the **assessment triangle** (i.e., cognition, observation, and interpretation), and, perhaps, the reemergence of performance assessment (e.g., Davey et al., 2015) as a means of assessing higher order thinking skills (e.g., Darling-Hammond & Adamson, 2010).

Policy developments also have played a role. For example, the *Race to the Top* grant application awarded points for "approaches for developing assessment items" and referred explicitly to ECD as an approach (Race to the Top Fund Assessment Program, 2010, p. 18181). With this incentive, the *Smarter Balanced*, *Partnership for Assessment of Readiness for College and Career* (PARCC), *Dynamic Learning Maps*, *National Consortium of State Collaboratives* (NCSC), and *ELPA21* multi-state assessment consortia all implemented versions of ECD. Moreover, ECD is mentioned regularly in conference papers, and is being taught in graduate courses. The word "evidence" is now prevalent in discussions about large scale assessment, which is a "significant shift from the traditional approach that was less specific in theory about how all the pieces of assessment link together, from task model to psychometric model" (J. Behrens, personal communication, July 28, 2015).

A practical matter that is rarely discussed outside of commercial testing vendors, test sponsors, and the media also has played a role: Vendors and their assessment program clients are highly motivated to find efficiencies and cost savings. For example, the unit

cost to develop a multiple choice item for operational use runs into hundreds of US dollars and more while costs for the development of essay **prompts** may reach as much as $15,000 (Bowie, 2015). The potential that principled design and development approaches could reduce the time and money spent editing and re-editing items, which we address later, may be influencing vendors to propose principled approaches in their responses to requests for proposals.

The "next generation" of content standards also plays a role. The kinds of assessments required to align with the rigorous demands of the *Common Core State Standards* and *Next Generation Science Standards* (NGSS) have led to demands for performance assessment, more constructed-response items, technology enhanced items, and new challenges to providing **accessibility** (AERA, APA, & NCME, 2014, chap. 3) for students with disabilities and English language learners. Principled approaches are likely to be particularly helpful for aligning these kinds of assessment activities with content targets and the cognitive demands intended by these standards. For example, a model of cognition, learning, or performance can explicate these targets and demands where the explicitness of broadly written content standards fall short.

Finally, principled approaches are in use in large-scale educational testing, but not widely so and perhaps with low implementation fidelity. We are not aware of information on fidelity when principled approaches are implemented. We do know that seasoned test developers do not find it easy to break out of their familiar, efficient, and deeply rooted cognitive routines for item development. Experience indicates that they may find the conceptual and procedural requirements of principled approaches difficult to penetrate and to implement (e.g., Hain & Piper, 2016). Likewise, we are not aware of efficacy studies that focus on the impact of principled approaches on test item and quality, efficiency, and development costs.

Overviews of Principled Approaches

In this section we provide overviews of the five approaches to principled design, development, and implementation and how each addresses the six elements of principled approaches: clearly defined assessment targets; statement of intended score interpretations and uses; models of cognition, learning, or performance; aligned measurement models and reporting scales; manipulation of assessment activities; and ongoing accumulation of evidence to support validity arguments.

Evidence-centered Design (ECD)

ECD is a framework for identifying, developing, and operationalizing theories and models of learning and cognition in assessment design and development. It makes explicit the assessment argument (e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006, Table 4.1) in the form of claims about what examinees know and can do based on evidence generated in the assessment process. The ECD process is organized in five layers. During the design, development, and implementation planning process, assessment designers cycle through these layers rather than

move through the layers sequentially (Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 2003). ECD may be the most widely implemented of the principled frameworks and most widely recognized.

In the first layer, **domain analysis**, assessment designers gather information about the domain of interest that might be useful for assessment design and development, including models or **theories of learning**, models of performance, specialized vocabulary, and the kinds of technology and tools used in the domain. In the second layer, **domain modeling**, assessment designers organize the information gathered during the process of domain analysis in a design document to support later assessment design and development decisions. **Design pattern** tools are used in ECD to help document and organize this information (e.g., Mislevy & Haertel, 2006, Table 2). A design pattern is a table with fields that prompt the assessment designer to record the **knowledge, skills, and abilities** (KSAs), the important content, and the important performances, among other things, that assessment development should include to support the development of a family of assessment activities. This information is made more specific in the next layer, the **conceptual assessment framework**.

In the third layer, the conceptual assessment framework, assessment designers create three model architectures: **student model(s)**, **task model(s)**, and **evidence model(s)**. These components further refine the information gathered and organized in the design document. The student model delineates aspects of the **targets of inference** that the assessment designer intends to make inferences about, given the purpose of assessment. Task models represent the content to be used to elicit student performance that will be used as evidence about the targets of inference in the student model. The content is described in terms of features that may be classified as characteristic, variable, or irrelevant. These content features would have been identified earlier in the process during domain analysis and may be based on research findings, expert judgment, or may be untested assumptions.

Each task model is used to generate multiple assessment activities that are explicitly related via their content features and, potentially, with similar psychometric features. The evidence model represents instructions for interpreting students' performance and consists of three parts: **work product** specifications, **evidence rules**, and the **statistical model**.

Work product specifications describe the structure and format of the performance that will be captured, evidence rules describe how to code work products (e.g., using a rubric for students' use of argument in science) to capture aspects of the construct, and the statistical model describes how the coding of the responses will be aggregated to make inferences about what students know and can do.

Layers four and five in ECD are *assessment implementation* and *assessment administration*, respectively. During the assessment implementation process the tools created in the conceptual assessment framework are used to write items and tasks, construct rubrics or other evaluation rules, and scale the assessment. During assessment administration, the assessment is administered and results are analyzed and reported; these practical implementation aspects are described in what is known as the four-process model.

ECD was implemented for the *PARCC* (see http://parcconline.org/), *Smarter Balanced* (see http://www.smarterbalanced.org/smarter-balanced-assessments/), NCSC

(see http://ncscpartners.org/Media/Default/PDFs/NCSC-Policymaker-Handout-2-20-14.pdf), and *Dynamic Learning Maps* (see http://dynamiclearningmaps.org/content/test-development) statewide assessment programs required under *Race to the Top*. In addition, *SRI International* supports other organizations' use of ECD (see http://www.sri.com/work/projects/padi-applying-evidence-centered-design-large-scale-science-assessment) while *Cisco Systems* (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004), the *Educational Testing Service*, the *College Board* (Huff & Plake, 2010), and other groups or organizations have implemented ECD for their own assessment initiatives.

Cognitive Design Systems (CDS)

The CDS approach (e.g., Embretson, 1998; Embretson & Gorin, 2001) centralizes the role of cognitive theory in assessment design and item development and validation. It includes both a *conceptual framework* and a *procedural framework*. The conceptual framework identifies and distinguishes two aspects of **construct validity**: **construct representation**, which corresponds to construct meaning, and **nomothetic span**, which corresponds to construct significance or utility (see also Kane, 2006, p. 46). Creating the conceptual framework for a test builds cognitive models and validation studies into the test design process and provides feedback to guide item development before items are tested with examinees. Construct representation specifically is "the processes, strategies, and knowledge structures that are involved in item solving" (Embretson, 1983; Embretson & Gorin, 2001, p. 349). Research in cognitive psychology is used to identify features of stimuli that can be manipulated to vary cognitive demands of assessment activities. Relevant studies often include statistical **item difficulty modeling** in order to identify such features. Nomothetic span "concerns the relationship of test scores to other measures" (Embretson & Gorin, 2001, p. 349). Evidence to support nomothetic span typically includes correlations of performance variables from the measure under development and external measures.

The procedural framework contains a series of steps to follow in order to integrate cognitive theories into test design. Steps in the procedural framework are presented sequentially (Embretson, 1998); however, the design and development process is iterative. The procedures guide item development and validation and relate examinee item solving processes to score interpretations.

1. *Specify goals of measurement* for both construct representation and nomothetic span
2. *Identify critical design features for tasks*, especially those features that can be manipulated to affect the processes, strategies, and knowledge required of examinees.
3. *Develop a cognitive model* to identify relevant processes, strategies, and knowledge, organized coherently, which requires a review of relevant research literature, operationalization of stimulus features that relate to complexity, and identification of the impact of these features on psychometric properties (i.e., empirical item difficulty and discrimination).
4. *Generate items by manipulating item stimulus features* to create items that are expected to vary cognitive process, strategy, and knowledge demands in intended ways.

5. *Empirically evaluate models for generated tests*, which means that the item generation system must be empirically evaluated, the cognitive model must be evaluated by successfully predicting item performance (e.g., response time and item difficulty) as a function of the stimulus features, and psychometric models are evaluated by their fit to the item response data; misfit may arise due to either convergent data (i.e., failure of relevant stimulus features to achieve hypothesized effects on item parameters) or divergent data (i.e., impact of non-construct relevant features on item parameters).
6. *Bank items by cognitive complexity* so that items automatically generated through this process are organized by their sources of cognitive complexity.
7. *Empirically validate* through studies of, for example, the relationship between examinee task solution processes and processes hypothesized in the cognitive theory.

The *Abstract Reasoning Test* of the *Armed Services Vocational Aptitude Battery* was constructed following the CDS approach (Embretson, 1999).

Assessment Engineering (AE)

The AE approach is a "highly structured and formalized manufacturing-engineering process" (Luecht, 2013, p. 3) with four stages: (1) **construct mapping** and **evidence modeling**, (2) task modeling, (3) designing item templates and writing items, and (4) calibrating items and quality control (see Luecht, 2013). The stages are designed and implemented to achieve "three fundamental assertions" (Luecht, 2013, p. 6), which are that (a) the content requirements and complexity of items differ across the examinee proficiency and test score scale, (b) a "family" of items can be designed from a model of task complexity that specifies declarative and procedural knowledge and other requirements for responding to items in the family, and (c) large numbers of items can be engineered within the same family with the same task complexity and psychometric (e.g., item difficulty) properties.

During the processes of construct mapping and evidence modeling, the assessment designer develops a **construct map**, which is a set of claims about examinees that are ordered along a complexity scale that coincides with the intended proficiency continuum and score reporting scale, similar to achievement level descriptors. During this stage, designers also create evidence models, or descriptions of what performance at each level of this ordered scale looks like. The second stage, task modeling, focuses on creating a set of specifications for a family of related task templates, which are themselves more detailed specifications for families of related items or assessment tasks. These specifications include detailed descriptions of the assessment targets, response demands of the items in the task family, as well as other item or task features that may impact cognitive complexity, and so relate back to both the construct map and the evidence models.

In AE, the specifications are written using a highly controlled language called a **task model grammar**. These grammars are potentially programmable specifications for generating items in the same family so that they are isomorphic in terms of cognitive complexity (i.e., declarative, procedural, and other response demands) and in location

on the proficiency scale. Once the task models are created, they can be arrayed along the complexity scale to create a task model map that portrays which locations along the proficiency scale will be given the greatest emphasis during task development.

Each task model is then implemented during the processes of designing item templates and writing items to develop item templates. The templates provide even more detailed specifications, including item format and scoring rules, manipulable features, and evaluation criteria. By systematically varying parameters within the manipulable features, item writers or programmed task model grammars can create multiple items from the same template. These items are expected to be co-located on the complexity/proficiency scale through their connection to the item templates and task models. During the final stage, calibrating items and quality control, items are field tested and calibrated using modern measurement models to confirm that the hypothesized complexity/proficiency ordering of items, templates, and task models actually holds, and to make adjustments where it does not.

The AE approach was used to demonstrate how to develop construct map versions of cognitive models (Gierl & Leighton, 2010), task models, and an associated task model map (Luecht, Dallas, & Steed, 2010), as well as to generate and evaluate 10,301 items based on 15 item templates (Lai, Gierl, & Alves, 2010) for the *Critical Reading and Mathematics* section of the *College Board's Preliminary SAT/National Merit Scholarship Qualifying Test* (PSAT/NMSQT).

The BEAR Assessment System (BAS)

The BAS is a construct modeling approach (National Research Council, 2001, 2006; Wilson, 2005) with four building blocks to guide assessment design and development; construct modeling is the process of creating working definitions of the assessment targets. Broadly speaking, the BAS describes a cycle of assessment that starts with a question; that is, an intended score interpretation and use that may be norm-referenced, criterion-referenced, or decision-based (Brown & Wilson, 2011).

The BAS uses construct modeling to describe the assessment target as a sequence of levels in the construct maps to illustrate the location of assessment tasks and examinees on an underlying scale of proficiency in relation to the construct. Formats of items and other assessment tasks are described in the stage of **item design**. Items and tasks are selected to elicit examinee knowledge, understandings, and skills hypothesized to match approximately to certain levels of the assessment constructs. Levels of responses to items and tasks, evaluated as levels of quality, are then described in the **outcome space**. The levels of quality of responses are illustrated using examples of examinee work to guide scoring and to correspond to levels of the construct maps. A measurement model is then selected to relate a data set based on the scores to levels of response quality for items and tasks using an item modeling approach. The building blocks are implemented as iterative steps in the assessment design process.

Construct modeling is undertaken to define a model of cognition as levels of proficiency along a continuous latent scale, and learning is conceived as progress from lower to higher levels of competence and sophistication (Brown & Wilson, 2011). In the BAS,

assessment tasks are selected to elicit evidence about examinee knowledge, understanding, and skills represented in the assessment construct and to reflect examinee progress along the learning continuum and latent scale. Assessment tasks that give rise to examinee performance and evidence of their knowledge in relation to the assessment construct, definitions in the outcome space, and the measurement model work together to implement the developmental perspective. Several such constructs may be under examination in a single test. Item modeling guides development of assessment tasks and the outcome space to elicit evidence of knowledge and skill in relation to the **target construct**. The BAS prescribes use of **Rasch models** for construct modeling and so that assessment tasks and examinee performance can be interpreted on **Wright maps** that show relevant parameters on a common scale (e.g., Wilson, 2005, pp. 85 ff.). This requires development of tasks that fit the rigorous assumptions of the Rasch model (e.g., approximately equal **item discrimination** across all items).

The BAS has been used in a range of applications, including classroom assessment (e.g., Kennedy, 2005), formative assessments embedded in an issues-oriented science curriculum for middle school grades (Wilson & Sloane, 2000), and an undergraduate assessment of conceptual understanding of scientific phenomena (Brown & Wilson, 2011).

Principled Design for Efficacy (PDE)

The PDE approach builds on ECD (Nichols, Ferrara, & Lai, 2016). As a result, it shares several concepts and tools such as domain analysis and domain modeling but emphasizes concepts and practices in unique ways. Specifically, the central role of KSA research from the learning sciences in construct definition and assessment activity design, as well as the emphasis on communication among stakeholders throughout the design and development process, stand out. The PDE approach is implemented as a principled enhancement to conventional, recognizable practices as illustrated in Figure 3.3 (discussed later) rather than a new, seemingly unfamiliar approach that can be off-putting to test developers and testing program managers.

The PDE approach to the design and development process consists of four stages and a framework with six design concepts. The four stages are named, designed, and carried out in ways that should be familiar and easily comprehensible to test developers and managers. During the first stage, *construct definition*, the assessment designer explores research literature from the learning sciences to define academic content standards or other assessment targets in terms of cognitive processes, knowledge structures, strategies, and mental models that are more fine-grained than educational content standards. The assessment designer uses research literature findings to describe features of stimuli and items that most effectively elicit the cognitive assessment targets, which are described as characteristic and variable content around which stimuli and items are developed.

During the second stage, *content creation*, assessment designers take advantage of the characteristic and variable content features to create stimuli and items that assess the full range of test-taker performance in relation to the assessment targets, as well as

rubrics for evaluating examinee test performance. The third stage, *generalization*, focuses on using the stimuli and items written during content creation activities to create reusable guidelines and specifications. Finally, during the fourth stage, *content re-creation*, content developers use the guidelines and specifications to generate additional numbers of stimuli, items, and rubrics.

The six design concepts for the work in the four stages are "intended to facilitate reasoning and communication in assessment design and development" (Nichols et al. 2016, p. 56). The *construct design concept* represents the assessment targets. The *evidence design concept* articulates features of test-taker responses that will be collected, as well as how they will be evaluated and aggregated. The *content design concept* specifies the features of stimuli and items that are needed to elicit those responses. The other design concepts include *communication with stakeholders* (e.g., examinees, item developers), *assessment implementation consistent with practical constraints*, and the *consequences* or **theory of action** for the assessment, which captures the intended outcomes of the assessment as well as the mechanisms for achieving them.

PDE has been used to develop a theory of action, task models, and performance assessment tasks for a system-wide elementary and middle school formative assessment program for the *Baltimore County (Maryland) Public Schools*, NGSS assessments for the Maryland statewide assessment program, and the *Insight Science and Dialogue for Language Learners* systems, two digital-device-based learning and formative feedback systems, now in development at *Pearson*.

Elements of Principled Approaches

The five approaches to principled assessment design, development, and implementation share five foundation elements that make them principled. These elements are organized under a sixth, organizing element, which is the primary goal of principled approaches: the ongoing accumulation and synthesis of evidence to build validity arguments to support intended interpretations and uses. We discuss each of the six elements in detail in the following sections.

As shown in Table 3.1, the foundation elements align with the assessment triangle in *Knowing What Students Know* (National Research Council, 2001, pp. 44–53 and Figure 2.1). Specifically, clearly defined assessment targets, statements of intended score interpretations and uses, and model of cognition, learning, or performance correspond with the *Cognition* vertex; aligned measurement models and reporting scales corresponds with the *Interpretation* vertex; and manipulation of assessment activities corresponds with the *Observation* vertex.

An "approach" is defined as a "way of dealing with something" (see https://www.google.com/#q=approach+definition). Principled approaches are thus not fixed formulas for achieving a desired outcome like a cookbook recipe; instead, they are more heuristic than algorithmic. Their concepts, procedures, and tools share features of medical diagnosis to guide treatment decisions, the process of evaluating and improving someone's tennis strokes, and training and curricula that guide teachers as they teach their students to read, write, understand mathematics, science, and social

Table 3.1 Foundation and organizing elements of principled approaches to assessment design, development, and implementation and their relationship to the assessment triangle.

Framework elements	*Assessment triangle alignment*
Organizing element	
Ongoing accumulation of evidence to support validity arguments	Overall evidentiary reasoning goal
Foundation elements	
Clearly defined assessment targets	Cognition
Statement of intended score interpretations and uses	Cognition
Model of cognition, learning, or performance	Cognition
Aligned measurement models and reporting scales	Interpretation
Manipulation of assessment activities to align with assessment targets and intended score interpretations and uses	Observation

studies, and to think. A "principle" is defined as "a fundamental truth or proposition that serves as the foundation for a system of belief or behavior or for a chain of reasoning" (see https://www.google.com/webhp?sourceid=chrome-instantandion=1andespv=2andie=UTF-8#q=definition%20principle). Principled approaches thus reflect this definition and recommend plans of action (i.e., the proposition) and practices (i.e., behaviors) that are expected to produce high quality, validly interpretable assessment information and form evidence into validity arguments (i.e., the chain of reasoning).

Overarching Organizing Element

The overarching organizing element that we describe in this section serves to guide thinking, planning, and decision making for the foundation elements: accumulating validity evidence and building validity arguments to support intended score interpretations and uses throughout the design, development, and implementation process. The organizing element relates directly to validity arguments, which provide "an overall evaluation of the claims in … the proposed interpretations and uses of the scores generated by the testing program" (Kane, 2013, pp. 10, 14). Validity argumentation and the ongoing collection and synthesis of evidence to support validity arguments are built into principled approaches and are explicitly part of design, development, and implementation processes.

The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) characterize sound validity arguments that integrate "various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (p. 21). Collecting evidence throughout the design, development, and implementation process enables test designers and developers to evaluate whether each decision will produce evidence and develop a coherent account to support intended score interpretations and uses. Building coherent validity arguments is often neglected in testing program documentation (e.g., Ferrara & Lai, 2016). Accumulating evidence to support validity arguments as a

guiding principle for the design, development, and implementation process should correct this neglect. Table 3.1 shows five foundation elements whose instantiation in real practice is driven by the goals reflected by the overarching, organizing element; we now discuss each of these in turn.

Foundation Element 1 – Clearly Defined Assessment Targets

This is the starting place for principled approaches to assessment design, development, and implementation. The first step is to define the KSAs that will be assessed. That definition, in turn, facilitates making statements of intended and warranted interpretations and uses of test scores and dictates the types of assessment activities that can be included in a test. Assessment targets (e.g., Stiggins, 1994) can be defined (a) via construct definition (e.g., AERA, APA, & NCME, 2014, p. 11; Ferrara & DeMauro, 2006, p. 605; Haertel, 1985; Messick, 1994); (b) via a model of cognition, learning, or performance (e.g., Nichols, Kobrin, Lai, & Koepfler, 2016); (c) by selecting academic content standards that will be targeted in a state assessment program; or (d) a combination of the three. Clear, explicit definitions of assessment targets guide decisions throughout principled assessment design, development, and implementation. Continuous focus on assessment targets ensures that all subsequent design, development, and implementation decisions are consistent with and provide evidence to support claims and validity arguments about test score interpretations and uses. It requires assessment program developers to provide evidence based rationales for all decisions, where evidence to support decisions is available, and logical rationales where evidence is not available.

While it may not be explicit in our summaries or other descriptions of these approaches, defining cognitively and developmentally grounded assessment targets requires systematic reviews of the learning sciences and other literatures for empirical results from studies of cognition, learning, and performance in the targeted assessment domain. The implication is particularly significant for current practice in educational testing, where lists of academic content standards define the assessment targets, and **certification and licensure testing**, where lists of KSAs, identified in job analyses, define the assessment targets.

Many researchers before us have proposed defining assessment targets as achievement constructs (e.g., AERA, APA, & NCME, 2014, p. 11; Ferrara & DeMauro, 2006, p. 605; Haertel, 1985; Messick, 1994). They argue, and we agree, that construct definitions that hypothesize item responding processes and relationships among responding processes across multiple items "can provide a stronger foundation for test development and score interpretation" (Gorin, 2006, p. 22). Gorin also points out that the "generality of language" of state test content standards and **performance level descriptors** is a "significant limitation for test development and validation" (2006, p. 21). Principled approaches require and assist assessment designers to use analysis and research to translate lists of KSAs into representations of how examinees develop in relation to a construct (e.g., using **learning progressions**); how they perform in relation to a construct (e.g., using achievement level descriptors); or their status on the construct (e.g., mastery or non-mastery of a domain). Similarly, principled approaches

require and assist assessment developers to use research on learning and examinee processes for responding to items to ensure that assessment activities are aligned with assessment targets – or the representations of lists of skills and knowledge.

Decisions on design elements. As we indicated earlier, decisions about numbers and types of test items, testing time limits, mode of administration, examinee response scoring rules, rules for combining response scores into aggregate scores, and the content and format of reporting and other feedback are test design decisions. While these are familiar decisions in conventional test development practice, they often are made for practical reasons that may not align tightly with intentions about score interpretation and use, and typically require compromises. Assessment design is design under constraints after all. Consequently, decisions about numbers and types of test items often are made based on how much testing time can be tolerated by examinees and their advocates (e.g., educators and parents), scoring costs, score report turnaround requirements, and technology gaps between what is desired and what is available. Assessment designers work within these constraints to align the allotted assessment activities with facets of the target construct to maximize trustworthy inferences about examinee status, level of development on the construct, or quality of performance. Principled approaches provide tools such as design templates (e.g., task models in ECD) to specify assessment task features and requirements to align with cognitive targets as well as content and other targets.

Development decisions. Item developers typically are hired based on their content area expertise, experience working with examinees, and experience in item writing. Typically, they receive item writing assignments that specify numbers and types of items and the content targets for their assigned items and they are trained to understand these specifications and follow **universal design** principles (e.g., Zieky, 2016). We often rely on the expertise and experience of item writers to produce items that meet other crucial development requirements, such as cognitive and linguistic demands (e.g., Ferrara, Svetina, Skucha, & Davidson, 2011) and specific procedural, strategic, and other cognitive targets. Quite often, these requirements are not specified nor discussed in detail. It is likely the case that item writers, in turn, rely on deeply rooted, automated cognitive routines to develop items to align with specified content targets and, perhaps, to align in limited, unexplicated ways with those cognitive targets.

Scoring rubrics sometimes are developed by scoring experts, often after constructed-response items are developed by item writers. This approach probably is a result of siloed expertise and responsibilities rather than collaboration, and can result in misalignment between response requirements and scoring criteria. Similarly, reading passages and other stimuli that may accompany an item or item set (e.g., artwork, graphs and other visuals, video and audio clips) often are selected for or by the item writers without much explicit attention to the content and cognitive targets that the item writer is required to target. Principled approaches provide tools such as design templates to specify assessment task features and requirements to align with cognitive targets as well as content and other targets. Such tools codify design decisions so that everyone in the design and development process can make decisions consistent with previous and subsequent decisions.

Implementation decisions. When tests are administered, they no longer are in the control of designers and testing program managers. This means that test administrators play key roles in supporting valid interpretations and uses of test scores by conducting sound administrations and helping to protect test security. Once score reports are released to examinees, score interpretation and use leave the control of assessment program managers. Principles to guide implementation activities include test administration guidelines and requirements (e.g., test security protections), response scoring criteria, psychometric analysis specifications and procedures, score reporting content and formats, and support for appropriate interpretation and use of test scores and other information. Thus, despite all the care that assessment designers and assessment program managers take to enable and support valid interpretations and uses of test information, test score information can be misinterpreted and misused during implementation and use without principles of implementation for guidance.

The responsible parties and administration conditions are typically well specified, even though perhaps not controlled as tightly as is necessary to support intended score interpretations and uses. This seems to be true especially in educational testing where reports of test security violations and administration errors may signal pervasive problems (e.g., Fremer & Ferrara, 2013). Procedures and criteria for scoring constructed-response items from large scale assessments has matured into industry-wide standards that exist but are not well documented, except for specific tests and programs. Psychometric analyses and results are available in technical reports which often are readily accessible (cf. Ferrara & Lai, 2016, pp. 606–611). Research on score report contents and format suggests that guiding principles are now emerging (e.g., Zenisky & Hambleton, 2012), and research on communicating results is increasing our knowledge about the links between visualization and instructional decisions (Dhaliwal & DiCerbo, 2015) and ways to visualize changes in learner beliefs (Shute, Jeong, & Zapata-Rivera, 2016). Documents such as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014, Chapters 4 and 6) provide guidance for these matters.

Foundation Element 2 – Statement of Intended Score Interpretations and Uses

This principled element is not widely practiced in operational testing programs (Ferrara & Lai, 2016, p. 607) and only occasionally explicitly referenced in the descriptions of the principled approaches we review in this chapter. However, stating intended score interpretations and uses for an assessment should be a first design decision, made simultaneously with defining assessment targets (e.g., AERA, APA, & NCME, 2014, pp. 11, 76, and standard 4.1; Bejar, Braun, & Tannenbaum, 2007), as it provides precisely focused guidance for all subsequent design, development, and implementation decisions. One recommended way of stating intended score interpretations in the test design process is to write proficiency level descriptors as a first step, which can inform **standard setting** and alignment studies later in the development cycle (Bejar et al., 2007; Egan, Schneider, & Ferrara, 2012, pp. 82–83, 91–93).

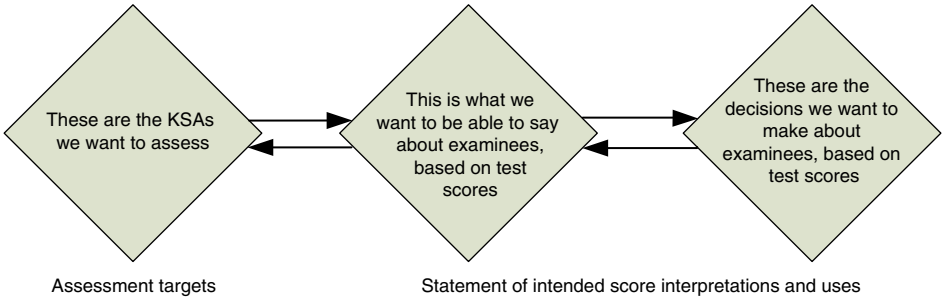

Figure 3.1 Logical and procedural relationship between two foundation elements in principled design: *specifying assessment targets* and *identifying intended test score interpretations and uses* simultaneously.

The logic of stating intended interpretations and uses in conjunction with identifying assessment targets can be portrayed graphically, as shown in Figure 3.1. Specifically, Figure 3.1 illustrates the way in which identifying assessment targets and stating intended score interpretations and uses are related logically and should be related procedurally. That is, sound decisions about examinees must be based on valid, evidence-supported inferences about examinee performance on the knowledge, skills, and abilities that are the assessment targets. The diamonds depict decision points and the arrows indicate the iterative nature of making such decisions. Once these foundation elements are decided, subsequent design, development, and implementation steps can be treated as a process of reasoning backwards, from implementation decisions, development decisions, and design decisions back to the original statement of intended score interpretations and uses.

Foundation Element 3 – Model of Cognition, Learning, or Performance

Principled approaches to design, development, and implementation include some form of a model of cognition, learning, or performance so that intended score interpretations and uses are connected to examinee thinking and achievement through test scores that result from scaling examinee responses to assessment activities.

Models of cognition. Leighton and Gierl (2007) define a cognitive model for the "broad problem-solving and skills" assessed on educational tests as a "simplified description of human problem solving on standardized educational tasks, which helps to characterize the knowledge and skills students at different levels of learning have acquired and to facilitate the explanation and prediction of students' performance" (p. 6). They also describe and evaluate the benefits of three types of cognitive models in educational testing: (a) the *model of test specifications*, in which content knowledge and skills are crossed in a matrix to guide test design and development; (b) the *model of domain mastery*, in which the entire set of skills that defines expertise or mastery of a circumscribed achievement domain guides test design and development; and (c) the infrequently implemented *model of task performance*, in which classes of assessment tasks

are generated and empirically validated to illustrate student thinking in solving educational tasks in a content domain. Similarly, as mentioned earlier, the widely cited assessment triangle in *Knowing What Students Know* (National Research Council, 2001) defines the *Cognition* vertex as "a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain" (p. 44). The *Cognition* vertex is used to identify a subset of "*targets of inference*" (p. 45; italics in original) that guide sampling from the larger theory of cognition to design a test of knowledge and skills.

Models of learning. Nichols, Kobrin, Lai, and Koepfler (2016) define domain-specific models of learning that "describe how learners acquire knowledge and skills and use them in different subject-matter areas." They offer examples of a *learning progression model* for student understanding of modeling in science, a *conceptual change model* for understanding rational number, and a *sociocultural model* for second language acquisition that is consistent with Maori language, culture, and ways of viewing the world. The report *Knowing What Students Know* (National Research Council, 2001) asserts that the "targets of inference [from test scores] should also be largely determined by a model of cognition and learning that describes how people represent knowledge and develop competence in the domain" (p. 178). The report proposes several features of models of learning to inform assessment design: they should (a) be based on empirical studies of learners in the target domain, (b) identify performances that distinguish beginning and more advanced learners in the domain, (c) describe the types of experiences that provoke learning, (d) convey the variety of ways that learners develop domain understanding and skill, (e) enable test designers to select portions of the *model of learning* to target in assessment design, and (f) be flexible enough so that learning and performance can be reported at fine grained and less detailed levels.

Models of performance. Reif (2008, Figure 2.1) proposes five requirements for achieving good intellectual performance generally, which also define good performances in particular intellectual domains specifically: usability, effectiveness, flexibility, efficiency, and reliability. In particular, good performances are "*usable* for accomplishing significant tasks [and] … should involve actual accomplishments, rather than mere talk" (p. 15; italics in original). And good performances are "*effective* in attaining desired goals" (p. 16; italics in original). Models of academic performance represent both the knowledge and skills that students have learned and the "form in which students' performances will be captured; i.e., the Work Product(s)" (Mislevy & Haertel, 2006, p. 10).

Foundation Element 4 – Aligned Measurement Models and Reporting Scales

The principal goal in selecting a measurement model is to provide psychometrics that align targeted models of cognition, learning, or performance and reporting scales with the intended score interpretations and intended uses. Measurement models provide

the means for scaling examinee test scores, task difficulties, and estimates of the magnitude of error in examinee scores. As Kolen (2006, p. 155) puts it:

> Scaling procedures are intended to facilitate the interpretation of test scores by incorporating normative, score precision, or content information into the score scales … [and they] are judged by how well they encourage accurate interpretations of test scores and discourage improper score interpretations.

The key concepts in this definition are to use scaling procedures that incorporate information into the score scales, specifically to support intended interpretations of the scores. The scaling process is achieved by implementing one or more measurement models that estimate one or more person parameters (i.e., a single estimate of examinee standing on a single construct or estimates of examinee standing on multiple constructs), item parameters (e.g., item difficulty and discrimination) or, in the case of **cognitive diagnostic models** (e.g., Rupp, Templin, & Henson, 2010), cognitive processes present or absent in an item's incidence matrix.

In principled approaches, the model of cognition, learning, or performance is the information that is incorporated into the scale scores to support appropriate interpretations and uses. Selection of measurement models that are aligned with the targeted model of cognition, learning, or performance is the principled element here. The panel on *Psychometrics for Performance Assessment* (Davey et al., 2015, p. 84) put it this way:

> Item response models link item response data to an underlying construct … [and] inform how estimates of the construct being assessed should be constructed from the item response data. [Thus,] a variety of models is needed to cover different types of evidence required to support score interpretations … [and] a variety of score scales may be used to describe standing on the underlying construct or constructs.

Put differently, measurement models provide an inferential bridge between examinee responses, aggregated across test items, and intended interpretations about examinee test performances and intended uses or actions based on those interpretations. These models relate examinee responses to the targeted construct and provide a means for determining examinee standing or current progress on the target construct. These formulations of scaling and psychometric modeling represent best practices in traditional psychometric decision making. They also provide a framework for selecting the right model or models given an assessment situation and its intended test score interpretations and use, cognitive, learning, and performance models, and score reporting plans. Gorin and Svetina (2011, Table 5.2) and Wise (in Davey et al., 2015, Table 5.1) summarize measurement models for a variety of types of response data and score reporting scales. Similarly, Almond, Mislevy, Steinberg, Yan, and Williamson (2015) describe the application of **Bayesian inference networks** to educational assessment design and analysis, especially for emerging, innovative designs, similar to Rupp et al. (2010) who do this for **diagnostic classification models**.

For example, scaling test data for unidimensional constructs that are targeted by items that elicit dichotomous response data can be achieved appropriately by widely used IRT models such as the Rasch or the three-parameter model or using

classical test theory approaches for creating test score scales. IRT models are available for response data for achievement constructs that require eliciting responses in polytomous, ordered categories (e.g., poor response = score level 1, partially acceptable response = score level 2, fully acceptable response = score level 3). *Multidimensional IRT models* are available, though not yet in wide operational use, for multi-faceted achievement constructs that represent distinguishable knowledge and skill dimensions. Other scaling procedures and measurement models are available to address dependency in responses (Yen & Fitzpatrick, 2006, pp. 123, 141–142), such as *testlet models*, and *hierarchical calibration models*, as prescribed in AE (Luecht, 2013). **Mastery tests**, along with a variety of cognitive diagnostic or diagnostic classification models (Rupp et al., 2010) are available to identify examinee mastery status on groups of discrete skills that have been identified explicitly in test item design.

It is common practice to decide, in a principled way, which measurement model(s) to use at the same time as deciding on numbers and types of items and testing time. Measurement models are selected often for pragmatic or philosophical reasons. For example, some people prefer the simplicity of one-parameter models, others the additional information about items from two- and three-parameter models. In practice, these decisions usually work out well because most educational achievement tests are designed to be essentially unidimensional and are limited to multiple choice and short constructed-response items that the standard IRT models support quite effectively. Other decisions may be less straightforward. For example, determining whether to scale a test with unidimensional models, whether to calibrate subscales of a test using separate applications of unidimensional models, or whether to scale using a multidimensional model requires analysis and judgment about the structure of the target assessment domain and its essential dimensionality and subsequent statistical testing of scored examinee responses (e.g., Yen & Fitzpatrick, 2006, pp. 123, 139).

Foundation Element 5 – Manipulation of Assessment Activities to Align with Assessment Targets and Intended Score Interpretations and Uses

Principled approaches to design, development, and implementation of assessments provide explicit procedures and tools to guide purposeful manipulation of assessment items and activities. The goal of this purposefulness is to align the responses that assessment activities elicit from examinees and the corresponding response evaluation criteria to the intended score interpretations and uses through the measurement and cognitive models. Features of individual items or assessment activities may be manipulated to achieve alignment with item cognitive complexity targets (e.g., Schneider, Huff, Egan, Gaines, & Ferrara, 2013), item difficulty targets (e.g., Ferrara et al., 2011; Gorin, 2006), or knowledge and skill requirements in each proficiency level descriptor (Ferrara et al., 2011; Schneider et al., 2013). More generally, work on **automated item generation** (Gierl & Haladyna, 2013) demonstrates the degree to which we are able currently to manipulate assessment activities to hit difficulty targets and align with proficiency level descriptors to support interpretations and uses. Furthermore, the rigor of expectations for accuracy, precision, completeness, and other requirements in response evaluation criteria can be manipulated to enhance this alignment.

Use of the Elements in Principled Approaches

The five principled approaches to design, development, and implementation of assessments reviewed in this chapter differ on some concepts and terminology, procedures, emphases, and other details. However, all five address the foundation and organizing elements that we have proposed; we now discuss the particulars of this conceptual alignment in this section.

Organizing Element – Principled Approaches as Process Models

Figure 3.2 portrays principled approaches to assessment design, development, and implementation as a process model, which brings to the forefront two important points. First, the organizing element, *accumulation and synthesis of evidence to build validity arguments to support intended interpretations and uses* appears in every step in the process of designing, developing, and operating a testing program. Second, only ECD, in the implementation and administration layers, and PDE, as part of its implementation concept, address test implementation explicitly. Next, Table 3.2 summarizes specifically how each approach addresses the five foundation elements.

Foundation Element 1 – Clearly Defined Assessment Targets

As the first column in Table 3.2 indicates, each approach defines assessment targets as a first step in designing and developing an assessment. It also is clear that each approach defines assessment targets in distinctive ways, following different processes.

Foundation Element 2 – Statement of Intended Score Interpretations and Uses

As the second column in Table 3.2 indicates, the main difference in these five approaches appears to be the extent to which these statements are explicitly required and the stage in which they are articulated. In ECD, the development of an evidence model provides

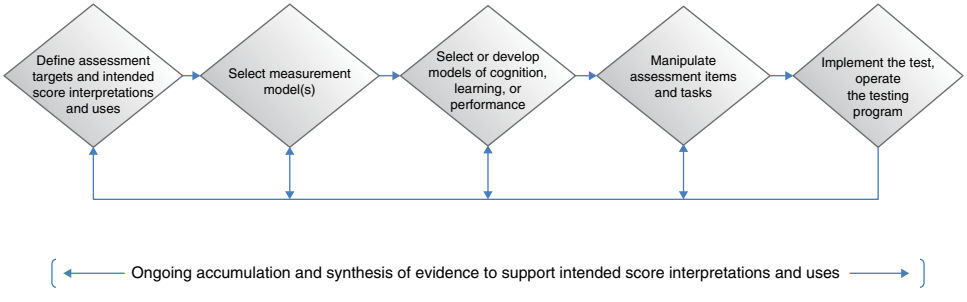

Figure 3.2 Process model for principled assessment design, development, and implementation to support intended interpretations and uses of test scores.

Table 3.2 Foundation elements of five principled assessment design, development, and implementation approaches.

Clearly defined assessment targets	*Statement of intended interpretations and uses*	*Model of cognition, learning, or performance*	*Aligned measurement models*	*Manipulation of assessment activities*
Evidence Centered Design				
Addressed in Domain Analysis, Domain Modeling and Conceptual Assessment Framework layers	Addressed as claims and creating assessment arguments in the Domain Modeling layer and in the Measurement Models used to aggregate data across assessment tasks, in the Conceptual Assessment Framework layer	Addressed in Domain Analysis and Domain Modeling layers and in delineating aspects of the target construct in the Student Model (in the Conceptual Assessment Framework layer)	Addressed in the Statistical Model, a component of the Evidence Model, which is in the Conceptual Assessment Framework layer	Addressed in Task Models, a component of the Conceptual Assessment Framework, supported by work completed in the Domain Analysis layer, and enacted in the Implementation layer
Cognitive Design Systems				
Addressed in Construct Representation in the Conceptual Framework and in step 1 in the Procedural Framework, *Specify goals of measurement*, as part of Construct Representation	Addressed as part of Specifying Goals of Measurement, in the Procedural Framework; measurement goals are required for construct representation and nomothetic span	Addressed in Construct Representation in the Conceptual Framework and in step 3 in the Procedural Framework, *Develop a cognitive model*	Addressed in the Procedural Framework, step 5, *Evaluate models for generated tests*	Addressed during research reviews to support Construct Representation in the Conceptual Framework and in the Procedural Framework, step 2, *Identify design features in the task design*, step 3, *Develop a cognitive model*, and step 5, *Evaluate models for generated tests*
Assessment Engineering				
Addressed at stage 1, Construct Mapping and Evidence Modeling; and at stage 4, Calibrating Items and Quality Control, when task models are adjusted to align with their intended ordering on the proficiency continuum	Addressed as a set of claims about examinees that are ordered along the proficiency continuum and score reporting scale, as part of Construct Mapping	Addressed in Construct Mapping and by specifying task model features that affect cognitive complexity, in Evidence Modeling	Addressed in stage 4, Calibrating Items and Quality Control, when hierarchical calibration places items and task templates on the examinee reporting scale, and in stage 2, Task Modeling, when task models are mapped to the complexity and proficiency scale	Addressed in stage 2, Task Modeling, and stage 3, Designing Item Templates and Writing Items

(*Continued*)

Table 3.2 (Continued)

Clearly defined assessment targets	*Statement of intended interpretations and uses*	*Model of cognition, learning, or performance*	*Aligned measurement models*	*Manipulation of assessment activities*
BEAR Assessment System Addressed in the Construct Modeling building block, when working definitions of assessment targets are created	Addressed at the start of the assessment cycle, with questions about intended score interpretations and uses	Construct Modeling defines levels of proficiency along a continuous latent scale	Rasch measurement models are used to relate scored responses to assessment tasks to levels on the proficiency scale, using Wright maps	Assessment tasks and the outcome space are developed to match proficiency levels, tested in observations and interviews, evaluated in Wright maps for match to proficiency levels and for model fit, and revised or rejected as necessary
Principled Design for Efficacy Addressed in stage 1, Construct Definition, and as part of the Construct design concept	Intended interpretations addressed by answering *What are you assessing?* Intended uses addressed by answering *What do you expect to happen when you assess?* and describing planned outcomes from using assessment information, as part of Consequences identification	Addressed as part of literature review in stage 1, Construct Definition, and as part of the Construct design concept	Addressed as part of the Evidence design concept	Addressed in stage 2, Content Creation, and stage 3, Generalization, and as part of the Content design concept

Note. All five approaches explicitly address ongoing collection and accumulation of evidence to support development of validity arguments.

instructions for interpreting examinee performance. In the CDS approach, interpretations and uses of scores can be stated as part of the specifying goals of measurement step within the procedural framework. In AE, construct mapping work involves developing a set of claims about examinees that are ordered along the construct-based proficiency continuum and score reporting scale in a way that is similar to achievement level descriptors. The BAS calls explicitly for identifying intended interpretations and uses by responding to questions. In PDE, target test content and a validity framework for claims about examinees and warrants for those claims are identified by answering the question such as "What are you assessing?" as part of the process of construct definition. Intended uses of test scores are stated by answering the question "What do you expect to happen when you assess?" and describing planned outcomes from using assessment information, as part of consequences identification.

Foundation Element 3 – Model of Cognition, Learning, or Performance

As the third column in Table 3.2 indicates, each of the principled approaches requires a model of cognition, learning, or performance. Most do not specify a type of model and each conceives of models in different ways. ECD does not prescribe specific types of models. Instead, the domain analysis process involves gathering information to model the assessment target domain and creating an assessment design and development framework using, for example, models of learning and performance, concepts, representational forms, terminology, technology, tools, and ways of interacting in the domain (Mislevy & Haertel, 2006). This analysis enables the process of domain modeling, which organizes knowledge and skills in the domain, their relationships, and corresponding assessment interpretation arguments, as well as the development of a conceptual assessment framework for the overall test design. In ECD, models may specify examinee cognition, learning, or performance in a domain, depending on what domain analysis and modeling yield, and they enable explicit validity arguments to support inferences about examinees.

The CDS approach similarly prescribes a conceptual framework to enable research on the validity of the assessment target construct and development of a research-based cognitive model to identify knowledge and **cognitive process** and **strategy** requirements or the target construct. The cognitive model identifies relevant processes, strategies, and knowledge and is organized coherently. Further, it prescribes conducting research to identify features of stimuli and assessment activities in order to vary their cognitive demands and psychometric difficulty. The model in CDS, then, appears to focus most explicitly on assessment task features and the complexity and difficulty of these features when examinees respond to assessment tasks.

The AE approach emphasizes assessment task features, complexity, and difficulty even more, with explicit recognition of their role in the psychometric scaling process. Assessment designers construct a construct map of claims about examinee knowledge and skills. These claims are ordered along a complexity scale that coincides with the intended proficiency continuum and score reporting scale. Assessment designers also create evidence models that describe performance at each level of the scale and cognitive task models to generate assessment activities that provide evidence of the

claims about examinee performance at each scale level. The cognitive model in AE is one of assessment task complexity at different levels of the test's proficiency scale, as illustrated in Luecht (2013, Figure 1).

The BAS defines cognitive models as levels of proficiency along a continuous latent scale and learning as progress from lower to higher levels of competence and sophistication. Finally, the PDE approach requires KSA research (Nichols et al., 2016, pp. 64, 78). This entails exploring learning sciences research literature to define academic content standards or other assessment targets in terms of cognitive processes, knowledge structures, strategies, and mental models that are more fine-grained than are educational content standards. The assessment designer uses research literature findings to describe features of stimuli and test items that most effectively elicit the assessment targets. Models in PDE may portray examinee cognition, learning, performance, or stages of learning leading to mastery of concepts or skills in a domain, depending on results yielded from KSA research.

Foundation Element 4 – Aligned Measurement Models and Reporting Scales

As the fourth column in Table 3.2 indicates, all five approaches require measurement models to calibrate items, create interchangeable test forms, and place examinee performance on a reporting scale that links evidence from examinee responses to assessment activities back to assessment targets and intended score interpretations and uses. ECD, CDS, and PDE require this linking but do not require specific measurement models. AE specifies hierarchical calibration models to place items and task templates on the examinee reporting scale. The BAS specifies Rasch models and use of Wright maps to ensure alignment of assessment tasks and the proficiency continuum.

Foundation Element 5 – Manipulation of Assessment Activities to Align with Assessment Targets and Intended Score Interpretations and Uses

As the fifth column in Table 3.2 indicates, all five approaches provide methods to manipulate features of assessment activities to elicit evidence from examinees to support intended score interpretations and uses. The primary tool in ECD for manipulating assessment activities is the task model. Task models describe the content knowledge and skill requirements that will be targeted in assessment activities. These requirements are characteristic (i.e., not manipulated), variable (i.e., manipulated to achieve more or less complexity, higher or lower task difficulty), or irrelevant (i.e., they should not influence examinee responses). The PDE approach uses similar tools, referred to more generally as item templates, and explicitly specifies that features of assessment activities should be manipulated to elicit needed evidence about the content and other features of accompanying stimuli (e.g., reading passages, visuals, audio, and video) that should be manipulated to enable the kinds of assessment activities and evidence intended.

The CDS approach specifies manipulation of assessment activity features to vary content and cognitive demands and requires validation that the manipulations result in items that are aligned with intended difficulty targets. It specifies a process for manipulating features; that is, collecting data on the relative impact of manipulated item features on item difficulty and discrimination. A goal of manipulating assessment activity features in the AE approach is to locate task templates and their families of items on prescribed levels of the proficiency scale and, thus, support intended interpretations and uses; the primary tool for accomplishing that is the task model grammar.

The BAS manipulates assessment activities to align with targeted levels of proficiency during item modeling and the definition of the outcome space, in small scale tryouts, and as part of evaluation of Rasch item modeling results.

Discussion

Relationship to Conventional Practices

The five principled approaches we reviewed in this chapter are not yet widely implemented in statewide summative testing programs or in licensure and certification testing programs. That is changing, however, as each approach has been used for various testing programs. One reason that adoption may be proceeding slowly is the challenge of internalizing complex terminology, concepts, and procedures (e.g., Hain & Piper, 2016) that seem unfamiliar compared to deeply ingrained concepts and practices of conventional test design, development, and implementation. In addition, because conventional practices have been successful for decades and efficiencies and procedural improvements have evolved, it may be difficult to recognize the enhancements that principled approaches build into the design, development, and implementation process. Finally, the validity argumentation approach to validation also is being adopted only slowly and is not widely practiced (see Ferrara & Lai, 2016). As test designers and developers and testing program managers adopt this mode of thinking, the need to adopt a principled approach to design, development, and implementation should become obvious.

Figure 3.3 illustrates conventional practice (white boxes) and how principled approaches enhance those practices (three grey boxes and grey background); both portrayals are simplifications of these very complex processes. In conventional practice, once a decision is made to develop a test, the test is designed primarily with test administration time, item type, cost, and other constraints in mind. Generally, those decisions are formalized in a blueprint and item development commences. After that, implementation begins, including field testing, development and application of scaling, equating, test-form assembly, score reporting, and score reporting procedures. We acknowledge that this bare-bones description obscures the complexity of these processes and the careful thought and rigorous execution required at each step.

From a practical point of view, principled approaches are not completely different from conventional practices and do not require learning completely new concepts and processes. Principled approaches require that additional thinking, additional steps and complexity, and documentation of design decisions and rationales – and more

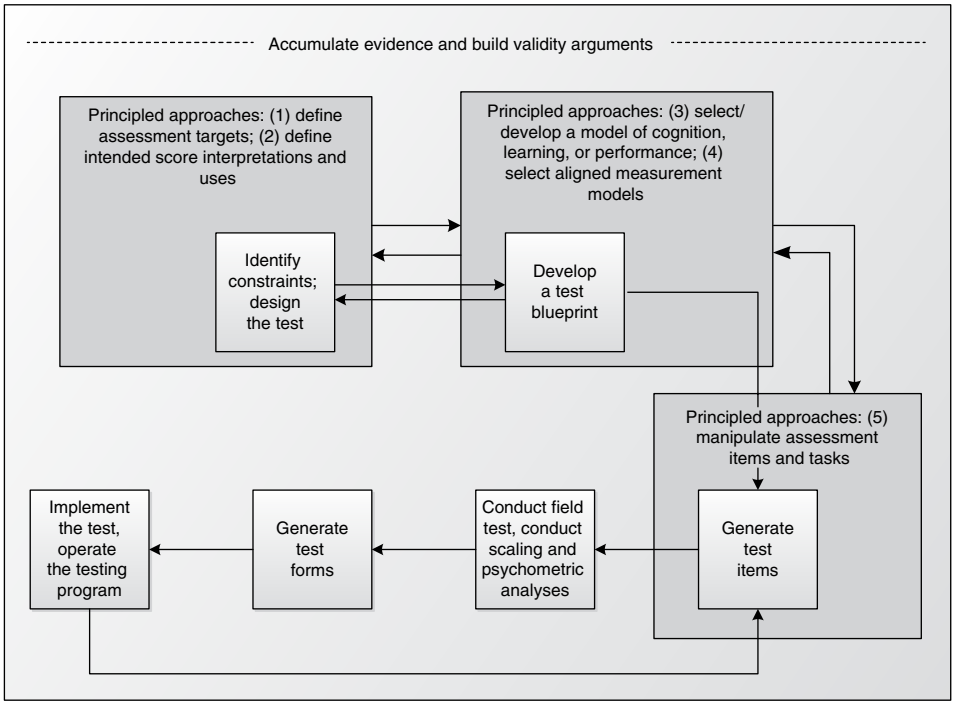

Figure 3.3 Conventional processes (white boxes) and processes based on principled approaches (foundation elements are numbered in the three boxes with grey background) for assessment design, development, and implementation showing overlap and differences.

work – be built into and around conventional practices. Furthermore, they require thinking about regular practices, especially regarding defining assessment targets, assessment activity design and development, and test score validation. Because of their generative properties, we can imagine using concepts as well as procedures and tools from these principled approaches to improve operations, item and test quality, and validity research for existing testing programs. For example, statements of intended score interpretations and uses can be made to correct an absence, or sharpened to improve clarity; cognitive or other models can be developed to explicate assessment targets and guide future item development; and task models can be developed to improve item quality and reduce attrition rates.

Furthermore, existing testing program operations can be enhanced to capture evidence throughout the process to support claims, interpretations, and uses of scores and to create validity arguments. For example, Huff and Plake (2010) edited a special issue of *Applied Measurement in Education* to illustrate how ECD was applied to the *College Board's Advanced Placement Program*. The articles in the special issue illustrate processes and tools, how knowledge and skills identified in domain analysis were translated into claims and evidence statements in domain modeling, how claims and evidence were used to develop proficiency level descriptors and aid standard setting, and development of task models and development of an overarching conceptual assessment framework.

Selecting among Principled Approaches

We suggest that ECD be the standard for comparison for all principled approaches. It appears to be the most widely implemented, its terminology and processes are highly evolved, and it is well documented. However, there are reasons that other approaches have been developed. The CDS approach has been used for design and development of mental rotation items, progressive matrix problems, object assembly items to measure spatial ability, nonverbal items to measure abstract reasoning, and mathematical word problems (Embretson & Gorin, 2001). Most of these are narrow constructs, and items for these tests have been developed and validated following highly constrained, rigorous procedures. It remains to be seen whether every step of the procedural framework and the level of rigor applied in these studies could be implemented for the broader achievement constructs, more stakeholder-inclusive processes, and challenging timelines required for state testing programs. The same may be true for licensure and certification tests, where stakeholders are involved perhaps to a lesser degree and where timelines may be slightly more manageable.

A prominent feature of implementations of the AE approach suggest that it may be particularly well suited to tests of learning and achievement for summative and formative interpretations and uses, especially for tests where it is necessary and feasible to generate large numbers of isomorphic items for large item banks. Construct mapping, evidence modeling, task modeling, and the use of hierarchical IRT modeling are focused clearly on developing assessment activities that support interpretations about examinee location on academic proficiency scales with proficiency level descriptors. In addition, the proposed use of task model grammars also supports this approach to assessment activity development. AE seems well suited for the design and development of tests of psychological constructs and licensure and for certification testing.

The BAS has been applied widely for classroom assessment, formative assessments embedded in a curriculum, and undergraduate education. We are not aware of its use for large-scale testing programs. The PDE approach has so far been used for the design and development of educational achievement tests, specifically statewide end of year summative tests. It was adapted from ECD for educational test designers and developers who work on statewide assessment programs. Its applicability is not limited to education. It seems readily adaptable to licensure and certification tests and psychological tests where target assessment constructs and assessment activities would benefit from application of learning sciences research results. One distinguishing feature of PDE is the consequences design concept. This concept requires explicating a theory or change – what is expected to happen as a result of interpreting and using test results – that clarifies the conditions for implementing a test that is required for intended outcomes to occur.

One might thus ask whether any of the five principled approaches seem particularly advantageous for particular item types or whether they merely accommodate people's innate desires to do things differently sometimes. One might ask whether some might be better for assessments with technology enhanced items, which so often are merely technology-enabled versions of **selected-response items** and short constructed-response items (e.g., Davey et al., 2015, Chapter II) or whether some are best for use for summative and high stakes testing purposes, as opposed to formative assessments?[2]

We do not see that as the case. Moreover, we do not think that principled approaches are practically useful only for new tests and assessment programs. Yet, it probably is the case that principled approaches are well suited for new programs that implement assessment standards that are unfamiliar to target examinees and users, where models and research can be fruitful for assessment activity design, score interpretation, and pursuing intended impacts (e.g., changes in teaching and learning, better job candidates).

Practical Challenges and Considerations

One of the chief enhancements that the five principled approaches can provide is explicating and tightening the chain of logic from conceiving the need for a test through its design, development, and implementation, so that evidence can be accumulated and intended score interpretations and uses can be supported in validity arguments. However, implementation of tests and assessment programs appear to be the weak link in this logic chain for at least two reasons.

First, tests leave the control of designers and testing program managers when they are administered. That means that test administrators play key roles in supporting valid interpretations and uses of test scores by conducting sound administrations and helping to protect test security. We probably do not do enough to train and to follow up when suspicions arise (e.g., Ferrara, 2014). In fact, responsibilities for testing programs are dispersed – but should be shared – among test designers, test developers, items writers, psychometricians, testing program managers, test administrators, and even policy makers. Perhaps implementation of principled approaches will highlight the interdependencies among these shared responsibilities. Second, as we discussed earlier, three of the five principled approaches do not refer to implementation – only ECD and PDE do in some form – so perhaps we should have referred to principled approaches to design and development only. But even in those two approaches there is no mention of test administrators, test security, and other practical considerations. Maybe that is not a surprise, though, because principled approaches are generally designed and developed for testing professionals, not the people we rely on for test administration and test security. It is a shortcoming that can easily be redressed.

We do not have systematically collected information on how often, under what conditions, and with what degrees and variations of implementation fidelity each of these five principled approaches have been used for operating tests and testing programs. Colleagues in educational testing have shared anecdotes that suggest two things. First, test development professionals find some of the terminology, concepts, and processes slow and cumbersome (see, for example, Hain & Piper, 2016, pp. 44–45). Second, they find some of the processes and tools burdensome to use; that is, they see implementation of these approaches for their existing practices merely as additional work with no apparent payoff to developers and the development process. We referred earlier to the functional utility goal of design engineering processes; another key concept in design that is relevant here is efficiency. Put differently, we advocate for the benefits of following principled approaches to design, development, and implementation because our

desired functional utility outcome, rigorously supported interpretations of test scores and subsequent decisions and actions, is a meaningful goal to strive toward. Test development professionals surely support that outcome and also desire efficiency.

Speculations on the Next Stages of Evolution

It seems likely that operational uses of principled approaches to design, development, and implementation will continue to grow, and to evolve in response to new developments in educational, psychological, and workplace testing. Partial or adapted implementations seem likely, as in the *College Board's Advanced Placement* example we mentioned earlier (Huff & Plake, 2010). Given the expanding role of learning sciences in assessment design, development, interpretation, and use, including intelligent tutoring, automated scoring, and cognitive diagnostic assessment models, we anticipate more widespread and diverse applications of principled approaches. For example, we expect to see principled approaches influence formative assessment programs and digital learning systems with embedded formative assessment activities. In addition, the rise of technology-enhanced assessment – particularly assessments embedded in learning games and technology-rich simulated environments (e.g., DiCerbo, Ferrara, & Lai, in press) – will almost surely prompt the use of principled approaches. Using interactions in these environments to construct measures of learning, cognition, and soft skills such as persistence are not yet well understood; principled approaches offer a means of avoiding construct irrelevance and mis-measurement. It also seems reasonable to expect that applications of principled approaches for technology-enhanced assessments might increasingly become associated with design-based research methods (Collins, Joseph, & Bielaczyc, 2004). These methods employ multiple iterations of design-test-revise cycles to build and resolve bugs in educational materials, a process not unlike the agile approaches favored by software developers.

We would like to see principled approaches evolve in a number of directions. Principled design is challenging and effortful. It is possible to go through the motions and carry out the steps of a principled process in a superficial way without engaging in the type of thinking that is necessary to realize the full benefits. For example, assessment designers who fail to dig deeply enough into construct definition can nevertheless apply tools and fill out templates for controlling item and task features. But failure to model sources of complexity in the assessment targets will undermine these efforts and may introduce **construct-irrelevant variance**. Thus, we hope to see more attention in the future to the implementation fidelity of principled approaches, with close attention to the cognitive aspects of principled design and development. In recognition of the challenges of understanding and applying principled approaches, we also would like to see better practical support for test and item developers who may be unfamiliar with the approaches. Organizations like SRI have experimented with the creation of interactive applications, such as automated item writing wizards, designed to reduce item writer cognitive load (Hamel & Schank, 2006). In Hamel and Schank's estimation, such tools serve the same function as widely available tax return software that populate cells in tax return forms on the basis of the filer's response to a series of questions.

Principled approaches also offer an opportunity to reduce test accessibility barriers to struggling learners (e.g., examinees with disabilities, English language learners). Much progress in this area is evident. For example, SRI has successfully integrated *Universal Design for Learning* (see http://www.udlcenter.org/aboutudl/udlguidelines) principles into the ECD process by representing accessibility impediments as additional KSAs in task models. Accessibility impediments are linked to characteristic item and task features in order to remove them as sources of construct irrelevance. Moreover, the NCSC assessment consortium (see http://www.ncscpartners.org/) employed ECD (Flowers et al., 2015) to address accessibility issues. They designed alternate assessments by defining assessment targets as alternative achievement standards tailored for students with severe cognitive disabilities. Despite this progress, published applications of principled approaches as a strategy for addressing accessibility issues are still relatively rare.

Finally, in order to support efforts at instilling principled approaches into day-to-day practices of assessment development, we see a need for evidence to support *our* claims about principled approaches (e.g., Brennan, 2010): (a) evidence of the efficiencies that are gained in using the tools and processes of principled approaches and, for that matter, evidence of improvements in validity, (b) improvements in clarity of terminology and concepts, usability of tools; and (c) improvements in efficiency of the additional activities and processes in principled approaches that are illustrated in Figure 3.3. The obligation to provide evidence suggests that we undertake evaluations of implementations of principled approaches and a research agenda on the benefits and outcomes we claim for them. The obligation to demonstrate improvements suggests that we might want to train test development and other professionals by starting not from the unfamiliar terminology and concepts of principled approaches, but from conventional test design and development practices to help them see how principled approaches enhance, rather than replace, those familiar recognizable processes.

For example, just as comparative studies of standard-setting methods in operational situations are rare (cf., Green, Trimble, & Lewis, 2003), probably because of the cost and upset to standard operations, using principled and conventional approaches simultaneously for a single operational assessment cycle may not be feasible. However, opportunities for naturally occurring experiments may already exist, for example, when a testing program shifts from conventional to a principled practice. In these situations, item quality based on expert reviews, item development efficiency, item revision and rejection rates, development costs, and item psychometric quality could be compared. Evidence from such studies would support the value claim for principled approaches and might lead to a broader and more principled adoption of these approaches. However principled approaches to assessment design, development, and implementation may evolve, it seems clear that operational practice is advancing toward them.

Acknowledgments

The authors thank Kristen Huff, Jackie Leighton, and André Rupp for their excellent insights and comments on earlier drafts of this chapter.

Notes

1 Principled approaches reflect definitions for the terms system, framework, and approach. A system is "a set of principles or procedures according to which something is done; an organized scheme or method" (see https://www.google.com/webhp?sourceid=chrome-instantandion=1andespv=2andie=UTF-8#q=system%20definition). A framework is "an essential supporting structure of a building, vehicle, or object" or "a basic structure underlying a system, concept, or text" (see https://www.google.com/webhp?sourceid=chrome-instantandion=1andespv=2andie=UTF-8#q=framework+definition). An approach is "a way of dealing with something" (see https://www.google.com/#q=approach+definition).

2 We distinguish externally provided formative assessments that may be available commercially, provided by a school district, state, or state assessment consortium or embedded in instructional materials from teacher classroom formative assessment practices.

References

Almond, R. G., Mislevy, R. J., Steinberg, L., Yan, D., & Williamson, D. (2015). *Bayesian networks in educational assessment*. New York, NY: Springer.

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.

Behrens, J. T, Mislevy, R. J., Bauer, M., Williamson, D. W., & Levy, R. (2004). Introduction to Evidence Centered Design and lessons learned from its application in a global elearning program. *International Journal of Testing*, *4*, 295–301.

Bejar, I. I., Braun, H. I., & Tannenbaum, R. J. (2007). A prospective, progressive, and predictive approach to standard setting. In R. Lissitz (Ed.), *Assessing and modeling cognitive development in school* (pp. 1–30). Maple Grove, MN: JAM Press.

Bormuth, J. R. (1970). *On the theory of achievement test items*. Chicago: The University of Chicago Press.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement*, *50*(1), 110–114.

Bowie, L. (2015, March 23). *Baltimore Sun*. Retrieved from http://www.baltimoresun.com/news/maryland/education/bs-md-test-cheating-20150322-story.html#page=1 (accessed July 30, 2015).

Brennan, R. L. (2010). Evidence-centered assessment design and the Advanced Placement Program: A psychometrician's perspective. *Applied Measurement in Education*, *23*(4), 392–400.

Brown, N. J. S., & Wilson, M. (2011). A model of cognition: The missing cornerstone of assessment. *Educational Psychology Review*, *23*, 221–234.

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *The Journal of the Learning Sciences*, *13*(1), 15–42.

Darling-Hammond, L., & Adamson, F. (2010). *Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education. Retrieved from https://scale.stanford.edu/system/files/beyond-basic-skills-role-performance-assessment-achieving-21st-century-standards-learning.pdf

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., & Wise, L. (2015). *Psychometric considerations for the next generation of performance assessment*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2015/jubf

Dhaliwal, T. & DiCerbo, K. E. (2015). *Presenting assessment data to inform instructional decisions*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

DiCerbo, K., Ferrara, S., & Lai, E. (in press). Principled design and development for embedding assessment for learning in games and simulations. In R. W. Lissitz & H. Jiao (Eds.). *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age Publishing.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance level descriptors: History, practice, and a proposed framework. In G. J. Cizek (Ed.) *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). New York, NY: Routledge.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*, 300–396.

Embretson, S. E. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika*, *64*, 407–433.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343–368.

Ferrara, S. (2006). Toward a psychology of large-scale educational achievement testing: Some features and capabilities (Editorial). *Educational Measurement: Issues and Practice*, *25*(4), 2–5.

Ferrara, S. (2014). Formative assessment and test security: The revised Standards are mostly fine; our practices are not (invited commentary). *Educational Measurement; Issues and Practice*, *33*(4), 25–28.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579–621). Westport, CT: American Council on Education/Praeger.

Ferrara, S., & Lai, E. (2016). Documentation to support test score interpretation and use. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 603–623). New York, NY: Routledge.

Ferrara, S., Svetina, D., Skucha, S., & Davidson, A. H. (2011). Test design with performance standards and achievement growth in mind. *Educational Measurement: Issues and Practice*, *30*(4), 3–15.

Flowers, C., Turner, C., Herrera, B., Towles-Reeves, L., Thurlow, M., Davidson, A., & Hagge, S. (2015). *Developing a large-scale assessment using components of Evidence-Centered Design: Did it work?* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Fremer, J. J., & Ferrara, S. (2013). Security in large scale, paper and pencil testing. In J. A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 17–37). New York, NY: Routledge.

Gierl, M. J., & Haladyna, T. M. (Eds.) (2013). *Automated item generation: Theory and practice*. New York, NY: Routledge.

Gierl, M. J., & Leighton, J. (2010). Developing construct maps to promote formative diagnostic inferences using Assessment Engineering. In R. Luecht (Organizer), *An application of assessment engineering to multidimensional diagnostic testing in an educational setting*. Invited symposium at the annual meeting of the National Council on Measurement in Education, Denver.

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice*, *22*(1), 22–32.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice*, *25*(4), 21–35.

Gorin, J. S., & Svetina, D. (2011). Test design with higher order cognition in mind. In G. Schraw and D. R. Robinson (Eds.), *Assessment of higher order thinking skills* (pp. 121–150). Charlotte, NC: Information Age Publishing.

Haertel, E. H. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, *55*(1), 23–46.

Hain, B. A., & Piper, C. A. (2016). PARCC as a case study in understanding the design of large-scale assessment in the era of the Common Core State Standards. In R. W. Lissitz & H. Jiao (Eds.), *The next generation of testing: Common core standards, smarter-balanced, PARCC, and the nationwide testing movement* (pp. 29–47). Charlotte, NC: Information Age Publishing.

Hamel, L., & Schank, P. (2006). *A wizard for PADI assessment design*. (PADI Technical Report 11). Menlo Park, CA: SRI International.

Huff, K., & Plake, B. (Eds) (2010). Evidence-centered assessment design in practice [Special issue]. *Applied Measurement in Education*, *23*(4).

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

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

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). New York, NY: Routledge.

Kennedy, K. A. (2005). The BEAR assessment system: A brief summary for the classroom context. Technical report no. 2005-03-01. Retrieved from http://bearcenter.berkeley.edu/bibliography/bear-assessment-system-brief-summary-classroom-context

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: American Council on Education/Praeger.

Lai, H., Gierl, M., & Alves, C. (2010). Generating items under the Assessment Engineering framework. In R. Luecht (Organizer), *An application of assessment engineering to multi-dimensional diagnostic testing in an educational setting*. Invited symposium at the annual meeting of the National Council on Measurement in Education, Denver.

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(3), 3–16.

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448.

Luecht, R. M. (2013). Assessment Engineering task model maps, task models and templates as a new way to develop and implement test specification. *Journal of Applied Testing Technology*, *14*. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/45254

Luecht, R., Dallas, A., & Steed, T. (2010). Developing Assessment Engineering task models: A new way to develop test specifications. In R. Luecht (Organizer), *An application of assessment engineering to multidimensional diagnostic testing in an educational setting*. Invited symposium at the annual meeting of the National Council on Measurement in Education, Denver.

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

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 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. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3–67.

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

National Research Council. (2006). *Systems for state science assessment*. M. R. Wilson and M. W. Bertenthal (Eds.). Washington, DC: National Academies Press.

Nichols, P. D., Ferrara, S., & Lai, E. (2016). Principled design for efficacy: Design and development for the next generation of assessments. In R. Lissitz & H. Jiao (Eds.), *The next generation of testing: Common core standards, smarter balanced, PARCC, and the nationwide testing movement* (pp. 49–81). Baltimore: Information Age Publishing.

Nichols, P. D., Kobrin, J. L., Lai, E. & Koepfler, J. (2016). The role of theories of learning and cognition in assessment design and development. In A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment*: Frameworks, methodologies, and applications. (pp. 15–40). Chichester, UK: John Wiley & Sons.

Race to the Top Fund Assessment Program, 75 Fed. Reg. 18,171 (April 9, 2010).

Reif, F. (2008). *Applying cognitive science to education: Thinking and learning in scientific and other complex domains*. Cambridge, MA: MIT Press.

Rupp, A. A., Templin, J., and Henson, R. A. (Eds). (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guildford.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual response demands, and item difficulty. *Educational Assessment*, *18*, 99–121.

Shute, V. J., Jeong, A. C., & Zapata-Rivera, D. (2016 in press). Visualizing the processes of change in learner beliefs. H. Jiao and R. W. Lissitz (Eds), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective*. Charlotte, NC: Information Age Publishing. Retrieved from http://myweb.fsu.edu/vshute/pdf/beliefs.pdf

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York, NY: American Council on Education/Macmillan.

Stiggins, R. J. (1994). *Student-centered classroom assessment*. New York, NY: Merrill.

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

Wilson, M., & Sloane, K. (2000). From principles to practice. An embedded assessment system. *Applied Measurement in Education*, *13*(2), 181–208.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: American Council on Education/Praeger.

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice*, *31*(2), 21–26.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa*, *20*, 79–87. Retrieved from http://www.sciencedirect.com/science/article/pii/S1135755X14000141

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 81–99). New York, NY: Routledge.