7

Designing and Developing Educational Assessments

Kristen Huff
Curriculum Associates

Paul Nichols
Retired

M. Christina Schneider
Cambium Assessment

Huff, K., Nichols, P., & Schneider, M. C. (2025). Designing and developing educational assessments. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 441–511). Oxford University Press.
DOI: 10.1093/oso/9780197654965.003.0007

The intended audiences for this chapter are students and professionals—both early career and seasoned—engaged in the design and development of educational assessments. The focus of this chapter is the design and development of large-scale interim and summative assessments of educational achievement for which the primary intended purposes are reliable classification of student performance into one of three or more categories, or performance levels, with the intended inferences of what students know and can do with regard to specified learning standards supported by a compelling validity argument. For the purposes of this chapter, we define interim assessments as those that occur multiple times throughout a school year and are used primarily to inform instruction and track student progress. The term summative is used to refer to assessments that occur less frequently, such as once per year, and are used primarily for educational accountability. We believe that the design of educational interim and summative assessments as defined here and with these respective purposes is the same.[1]

Since the rise and proliferation of federally mandated end-of-year testing began in 1994 (Improving America's Schools Act, 1994), the educational measurement field has learned a great deal about what is needed to design and develop assessments that are intended to support interpretations about what students know and can do (Ferrara & DeMauro, 2006; Lane et al., 2016; Mislevy, 2006; National Research Council [NRC], 2001; Pellegrino et al., 1999, 2016; Rupp & Leighton, 2017). However, many of the recommendations by these scholars and practitioners are still, frustratingly, emergent rather than commonplace (Ferrara & DeMauro, 2006; Huff et al., 2017). Supporting valid inferences about what students have learned requires more sophisticated approaches to test design and development than in the past, when most educational tests were designed primarily, if not solely, for various norm-referenced decision-making or rank ordering (e.g., college admissions, to select the top 2% for gifted and talented programs, or the bottom 20% for special services). We posit that contemporary educational testing is faced with additional complex challenges that routine design practices simply cannot meet; we group these challenges into four categories: (a) the complexity of target constructs, (b) using assessment results for multiple purposes, (c) assessment quality, and (d) accessibility for increasingly diverse student populations.

First, the constructs of interest are being articulated in more sophisticated ways than in the past, as our scholarship continues to evolve in the learning sciences and as learning standards begin to reflect the sociocognitive nature of learning: how students develop the kinds of deep understanding required not only for disciplinary practice but also that set the stage for additional learning (Mislevy, 2006; NRC, 2000, 2001, 2002, 2005; Penuel & Shepard, 2017). The following are examples of standards and frameworks that reflect increasing complexity than their respective prior versions: The Common Core State Standards (CCSS; National Governors Association Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010), the Next Generation Science Standards (NGSS; NGSS Lead States, 2013), the National Assessment of Educational Progress (NAEP) Technology and Engineering Literacy Framework (National Assessment Governing Board [NAGB], 2018); the 2026 NAEP

Mathematics and Reading Frameworks (NAGB, 2021a, 2021b) and the Advanced Placement curriculum frameworks for science and history (College Board, 2019).

For example, as of this writing, most states have based their K–12 reading learning standards on the CCSS, which set expectations for comprehension from multiple sources, such as reading Martin Luther King's "Letter from a Birmingham Jail" and listening to one of his speeches and then responding to a writing prompt by citing evidence from both sources. In mathematics, there is now an expectation that students go beyond procedural fluency to conceptually understand multiple ways to problem-solve as an aspect of demonstrating proficiency in mathematical practices. Another example is the multidimensional nature of the NGSS, which integrate concepts that span the science domains, discipline-specific knowledge, and scientific practices. Therefore, the tasks we devise to measure these complex learning standards with unprecedented cognitive load, the coding schemes we use to score student responses, the measurement models we use to estimate student proficiency, and the concepts and indices used to evaluate technical quality of the assessment all need to reflect the intended complexity of the construct of interest. As an example of the new demands on evaluating technical quality, the conventional notions of comparability and reliability will likely need to evolve to keep up with contemporary needs.

Second, in addition to the challenges of measuring student proficiency with regard to sophisticated constructs, test users are demanding that educational assessments, both interim and summative, serve a variety of purposes as all manner of stakeholders—policy makers, parents, teachers, and students themselves—call for less testing, which results in the need for a greater variety of uses of the assessments that remain in place (Hart et al., 2015; Huff & Goodman, 2007). In practice, the results of interim and summative assessments are used to inform determinations at various levels of the educational system, including but not limited to instructional next steps at the individual or group level, whether the student needs additional testing for English language proficiency or dyslexia, grade promotion, algebraic readiness, educator effectiveness, school ratings, and district- and state-level policy decisions about curriculum and professional development resources. According to the *2014 Standards for Educational and Psychological Testing* (see Chapter 1; American Educational Research Association [AERA] et al., 2014), it is incumbent on the user to collect and evaluate evidence to support the assertion that the inferences from the assessment are valid for any use that is additional to or a departure from the purposes and use for which the assessment was originally designed and, presumably, validated (see also Lane & Marion, this volume). To meet the demands of less testing but increased usage of the results of any single test, test providers are looking for ways to design assessments from the beginning to serve more than one purpose.

With increased public discourse about the role of educational testing in schools, a third challenge has emerged that is not likely to subside anytime soon: Test quality is under scrutiny. As a condition for some federal education funding, the United States Department of Education requires state summative assessments to be reviewed for

technical quality by a panel of peers. Similarly, some states are requiring reviews for interim assessments before they are approved for use in classrooms (e.g., Louisiana Department of Education, 2020; South Carolina Department of Education, n.d.). As the level of scrutiny increases, there is greater burden on the test makers to have documented, transparent, and understandable validation arguments that support the variety of intended inferences about what students know and can do.

The fourth and last category of contemporary assessment design challenges is the increased need to ensure that our assessments provide results that support valid inferences about all students regardless of their sociocultural background, including home language, or disability.[2] We refer to this aim as ensuring that our assessments are accessible to all students. Teaching and learning are fundamentally sociocultural endeavors (Penuel & Shepard, 2017) that need to acknowledge and build on students' background knowledge and cultural experience. Assessment needs to reflect this principle to remain authentic and relevant to students' classroom experience. Additionally, there is a growing research base exploring the hypothesis that without this contextual relevance, we will not engage students during the assessment, which will demotivate them and thus undermine the accuracy of the resulting scores as a measurement of best performance, in part because of lack of consideration of construct-irrelevant factors that are present when sociocultural factors are not considered (Brown et al., 1989; Wise, 2020).

Students with disabilities must also have equal opportunity to demonstrate what they know and can do through accessible assessments. Implementation of the universal design for learning principles that guide assessment design (CAST, 2018; Johnstone et al., 2006) have been available to inform practice for some time for both paper-and-pencil and computer-based assessments. In addition, as more assessment systems are delivered via the Internet, many providers hold themselves accountable to the Web Content Accessibility Guidelines (WCAG; World Wide Web Consortium (W3C), 2018), which address features such as use of color and images, speech-to-text availability, captioning, and more. Simultaneously ensuring that students from various sociocultural backgrounds are engaged during the assessment, that the assessment is free of barriers for students with disabilities and multilanguage learners, and that the target of measurement has not been compromised requires a more contemporary approach to assessment design than is currently commonplace. When the targets of measurement are defined such that all ambiguation is eliminated, then discussions about what constitutes construct-relevant versus construct-irrelevant variance can occur with a level of precision that is typically absent, allowing for designing assessments that eliminate barriers to access that might otherwise be present. Creating assessments that are free of bias, are accessible to all students, and are culturally and linguistically responsive requires undisputed clarity on the target of measurement by the full interdisciplinary team.

It is in the context of these contemporary challenges that we write this chapter. To meet contemporary demands, educational achievement tests must be designed with more deliberate, interdisciplinary decision-making about each design feature and must make all assumptions and rationales transparent for interrogation. The intended inferences

about what students know and can do must have a compelling—and understandable—validation argument that resonates with policy makers, educators, and parents. Transparent design and a compelling, understandable validation argument are critical to support the varied intended purposes and uses of educational interim and summative assessments, especially in a climate when educational testing is under unprecedented scrutiny and criticism. In this chapter, we illustrate how principled assessment design (PAD) is not a rigid set of processes that require a new jargon-filled lexicon, but rather a discipline requiring a particular mindset and the use of tools that can help us build assessments that meet contemporary educational needs.

Two assertions are made throughout the chapter. First, assessments designed with conventional approaches do not adequately serve the purposes for which assessments are used in 21st-century educational systems; PAD offers a solution. Given the primary use case for large-scale assessments of educational achievement—to support valid inferences of what students know and can do in relation to a given set of learning outcomes for a given purpose and use—we, as a field, can no longer support approaches to educational assessment design and development that implicitly assume the measurement of a largely fixed latent trait for the purpose of rank ordering and selecting students and that define adequate validation evidence as a disconnected series of post hoc analyses to be collected and published in the technical manual. This is especially true as the constructs to be assessed become more complex, assessment results are used for a variety of purposes, assessment quality is under more public scrutiny, and our understanding of accessibility becomes simultaneously broader and more nuanced.

Conventional approaches to educational assessment design are fragmented and must be integrated and supplemented to best serve the intended role of large-scale educational assessments, whether designed to inform instruction, to make student-level decisions, or for broader accountability purposes. Fragmentation in design and development undermines the coherence required to support the inferences the assessment results are intended to support within the larger context of the educational endeavor (Herman, 2010; Huff et al., 2017). Thus, our second assertion is that PAD integrates what is fragmented in conventional approaches into a coherent design approach and validation argument. For example, historically there has been a stark divide in teams and processes between the item writing and the analysis of the resulting data, as if the creation of the test were an assembly line or a relay with handoffs. One clear consequence of this disjunction is the existence of test content specifications that are separate from the test psychometric specifications. In conventional approaches to assessment design, these two design elements are developed separately by respective teams with the exception of a few touchpoints. This divide in mindset and practice can be bridged through the discipline and practice of PAD.

PAD is an umbrella of test design approaches including construct-centered measurement (Messick, 1994; Wilson, 2005), cognitive design systems (Embretson, 1998), evidence-centered design (ECD; Huff et al., 2010; Mislevy & Haertel, 2006; Mislevy et al., 2003; Pearlman, 2008a, 2008b), principled design for efficacy (Nichols et al., 2016),

and assessment engineering (Luecht, 2013). PAD is distinguished from conventional approaches to assessment design and development by three overarching characteristics (see Figure 7.1).

First, the construct is defined by how students learn and build knowledge in the domain of interest, with a careful consideration that students from different cultural backgrounds bring different funds of knowledge that need to be valued and serve as the foundation for learning (Gay, 2000; Randall, 2021). A compelling example of construct definition is how the Advanced Placement (AP) program redesigned its science and history exams over the course of several years beginning in 2005 in response to criticism from the NRC (2002), where both the AP and International Baccalaureate examinations were criticized for overemphasizing declarative and procedural knowledge and for a lack of pedagogical and learning science research in science and historical thinking to inform assessment design. Given that assessments have consequences for what is taught and valued in the classroom, the NRC challenged the test providers to do better at ensuring that what is measured on the exam is also valued in the discipline. Other examples of this move away from overvaluing declarative and procedural knowledge as targets of measurement are the focus on scientific practices in the NGSS and the focus on mathematical practices in the CCSS. Both sets of learning standards were informed by research from the learning sciences and practitioners in the respective disciplines (CCSSO, 2010; NGSS Lead States, 2013).

A second distinguishing characteristic of PAD is the explicit articulation of all assumptions, design decisions, and rationales for those decisions, with particular attention to our need to serve a diverse student body—students from various sociocultural backgrounds, multilanguage learners, and student with disabilities. The various articulations result in a set of design tools that are used throughout the design and development process and are also subject to continuous improvement throughout the life of the assessment program as data are collected from students and used as feedback to inform refinements to our assumptions, design decisions, and tools. Use of these

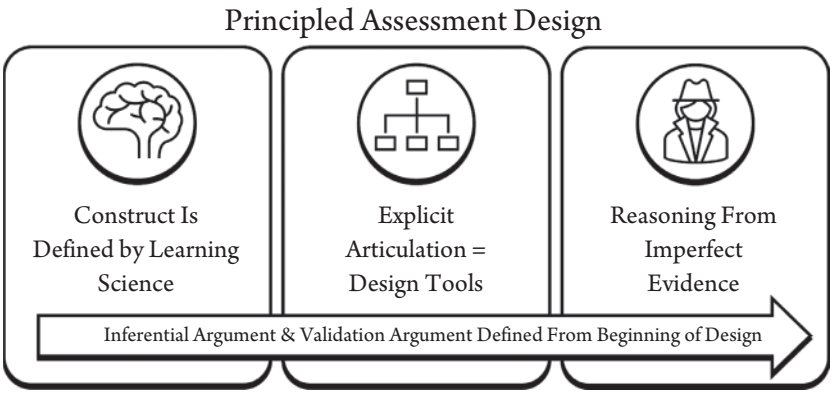

FIGURE 7.1
Distinguishing Characteristics of Principled Assessment Design

tools consistently by all players in the design and development endeavor not only helps to ensure coherence from construct definition through task development and score inference (and all the steps in between), but also serves as infrastructure for the inferential, evidential, and validation arguments. Briefly, the inferential argument is the chain of reasoning that is woven throughout the assessment design (e.g., How do these items measure the intended construct and support the claims about what students know and can do?), the evidential argument is the evidence for the inferential argument, and the validation argument is how well the evidence supports the intended inferences.

Finally, the last distinguishing characteristic of PAD is an adherence to the mindset that assessment is the process of reasoning from imperfect evidence, and that process begins with design. Reasoning from first principles—in short, leaving no assumption left implicit—is an explicit requirement throughout the design, development, and evaluation process.

The benefits of this shared discipline of PAD are many. Employing PAD will force us to make intentional and explicit the necessary relationships among performance-level descriptors (PLDs), task design, and scale properties that result in a strong body of evidence to support inferences about what students know and can do. In other words, we need the psychometric properties of the scale (e.g., sufficient measurement information across the scale, especially at various cut scores) to support reliable classification of students into performance categories and for those classifications to support valid inferences about what students know and can do. Rather than leaving this to happenstance, we need to purposefully design tasks that require students to demonstrate the knowledge and skills associated with various performance categories. PLDs are our guidance for task design at the beginning of assessment design and our guidance for score interpretation once we have scores for each student. That is, the PLDs articulate the inferences we eventually want to make about what students know and can do, and using them as a primary element of task design is a strong foundation for the inferential and validation arguments for the assessment.

In PAD, there is an almost obsessive focus on clarity, articulation, and documentation of all aspects of assessment design, especially the targets of measurement, the evidence required to support inferences about those targets of measurement, and the tasks that are best suited to collect that evidence given the constraints of the assessment. Later in the chapter, we will address how the evidentiary arguments supporting how specific tasks yield valid evidence for inferences regarding specific targets of measurement are a subcomponent of a larger evidentiary argument for the assessment as a whole. Suffice to say for now that these intentionally designed layers of evidence, made explicit and transparent, result in increased clarity in the field about the intended targets of measurement of the assessment and the constructs to which the inferences about students can be generalized, which results in greater coherence between what inferences the assessment results support with regard to what students know and can do and the role the assessment plays in a larger educational context. This clarity and coherence help the assessment results better serve the diverse needs of policy makers, educators, students, and parents—and it is our hope that this, in turn, will make the results more meaningful

for them. As a field, we must make our assessments more meaningful and useful for educators, parents, and students—not only because it is the right thing to do to support teaching and learning, but also in light of the many demands facing educational testing in the early 21st century. PAD gives us a way to do that.

SIMILARITIES TO CONVENTIONAL ASSESSMENT DESIGN AND DEVELOPMENT

In 1989, Jason Millman and Jennifer Greene wrote the test development chapter for *Educational Measurement*, third edition, which opened,

> This chapter is about making tests. It is directed to the professional test constructor, not to the classroom teacher. Our goal is to emphasize options for specifying and developing tests, not to produce a procedural manual. (p. 335)

We have the same goal. There are many sources that outline in much detail the precise steps for developing an assessment. Specifically, the *Handbook of Test Development*, second edition (Lane et al., 2016), includes 32 chapters devoted to every aspect of test development, administration, scoring, and evaluating validity evidence for the intended purpose and use of the assessment. The editors did an excellent job of balancing conventional approaches and PAD in their selection of authors and chapter topics. Another excellent resource is the 2013 publication by CCSSO and the Association of Test Publishers entitled *Operational Best Practices for Statewide Large-Scale Assessment Programs*.

Operationalizing PAD does not exempt the test developer from the nuts and bolts of test development; appropriate subject matter experts (e.g., classroom teachers, learning scientists) are engaged in the process, items are developed and reviewed against various criteria, items are field tested, and psychometric analyses are conducted (including but not limited to analysis of item difficulty, discrimination, differential item functioning, scaling, equating, standard setting, and more). Rather than devote this chapter to an explication of these well-documented processes and procedures, we have attempted—like many before us since the early 1980s—to question the status quo and make a compelling argument for shifting our practice to better support inferences about what students know and can do. As Huff et al. (2010) said of ECD—which applies equally to PAD—"ECD can be a first step toward challenging the assumptions and 'breaking out of the current paradigm' of large-scale assessment" (p. 316).

THE EVOLUTION OF PAD

There is a saying, "Nothing comes from nowhere." This is true of PAD. The current PAD approach grew from a number of influences, some obviously traced but others more obscure. In this section, we will review the following influences on the development of PAD: the integration of psychology and psychometrics, the focus on evidential reasoning, and the influence of design science.

Integration of Psychology and Psychometrics

The writers and researchers advocating close integration of psychology and psychometrics have been an important influence on the development of PAD. Leveraging contemporary research on how students learn and build knowledge to define constructs distinguishes PAD from conventional approaches to test design. Early advocates often approached the integration of psychology and psychometrics from the perspective of validity (Anastasi, 1967, 1986; Lindquist, 1951; Loevinger, 1957; see Snow and Lohman, 1989, and Lawrence and Shea, 2008, for more extensive reviews). For example, Loevinger (1957) made the pithy observation about test development driven by criterion-related validity (here, *criterion* is used to mean what the test predicts well):

> The argument against classical criterion-related psychometrics is thus twofold: it contributes no more to the science of psychology than rules for boiling an egg contribute to the science of chemistry. And the number of genuine egg-boiling decisions which clinicians and psychotechnologists face is small compared with the number of situations where a deeper knowledge of psychological theory would be helpful. (p. 82)

The 1980s saw a proliferation of authors both urging and demonstrating a closer coordination of psychology and psychometrics. An impetus for this movement may have been that psychology was escaping from the restrictions of behaviorism (Lachman et al., 1979). As Glaser (1981) noted, learning scientists had begun studying learning and individual differences in domains of interest to educators such as mathematics and reading, whereas past psychological research concentrated on the discovery of domain-agnostic mechanisms of thinking and learning. By the end of the decade, progress in the integration of learning science and psychometrics led Snow and Lohman (1989) to question the usefulness of conventional, trait-based, psychometric score interpretations:

> The implication is that sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst, they have misled scientists, and the public, into thinking of fundamental, fixed entities, measured in amounts. Whatever their practical value as summaries, for selection, classification, certification, or program evaluation, the cognitive psychological view is that such interpretations no longer suffice as scientific explanations of aptitude and achievement constructs. (p. 317)

The 1990s saw the promise of integrating psychology and psychometrics turned into practical applications. Researchers proposing practical applications of an integrated approach included Lohman and Ippel (1993), who described a cognitive diagnostic framework for creating assessments that took advantage of research identifying item features that could be manipulated to vary cognitive complexity and item difficulty. Another researcher proposing an integrated approach was Nichols (1994), who claimed that theories of learning and cognition were well suited to informing assessment design

and development. Nichols proposed a framework for test development that consisted of an observation design and a measurement design. The observation design describes the characteristics of assessment activities, such as tasks or items, that make demands on the test taker, how these characteristics are to be organized in the construction and ordering of tasks or items, and the nature of the responses required. The measurement design defines the object of measurement and describes the procedures to assign a value or category to an object of measurement. In addition, Embretson (1998) described a cognitive design system consisting of a conceptual framework that integrated models of learning and cognition into test design and a procedural framework that described the steps necessary to implement the conceptual framework.

Focus on Evidential Reasoning

Another important influence on the development of PAD has been writers, particularly Schum (1994, 2009) and Toulmin (1958, 2006), who examined the evidence we gather and use as a basis for making claims in the fields of law and philosophy and that have been applied to assessment design beginning with ECD.

From Schum (1994) came an emphasis on first principles that encouraged PAD practitioners to step outside the implicit assumptions and customary practices of conventional psychometrics and reconsider the evidentiary reasoning underlying probabilistic inferences about emerging complex constructs. When reconsidering conventional psychometrics, PAD practitioners have borrowed from Schum the principles of relevance, credibility, and force of evidence. First, questioning the relevance of evidence has motivated practitioners to critically examine the bearing of evidence on intended score interpretation and use leading to, primarily, a greater inclusion of cognitive process evidence and less reliance on correlational and predictive evidence. Second, questioning the credibility of evidence has resulted in practitioners articulating and carefully examining the trait-based assumptions implicit in conventional psychometric practices. Finally, a skeptical stance toward the relevance and credibility of conventional psychometric evidence has led practitioners to reconsider the weight given to particular evidence in a score interpretation and use argument made within a principled assessment framework.

From Toulmin (1958, 2006) came an emphasis on practical arguments, such as those used in nonmathematical fields like law, rather than formal logic. When arguments are evaluated in formal logic, the goal is to determine whether an argument is true or false. Practical arguments cannot be usefully evaluated by the rules of formal logic because the assumptions used in practical arguments cannot be taken for granted; the available evidence is often incomplete or questionable. As such, the practical argument is, at best, convincing or plausible rather than true or false.

The combined influence of Schum (1994) and Toulmin (1958, 2006) contributed to the role of evidentiary reasoning as an integral part of PAD. The emphasis on making evidential reasoning visible through the explicit articulation of all assumptions, design decisions, and rationales for those decisions is a distinguishing characteristic of PAD.

Similar influences are apparent in the development of an argument-based approach to validity (Kane, 2006, 2013; Messick, 1989) and design propositions in design science (van Aken & Romme, 2009), creating compatibility with PAD.

Design Science

The emergence of design science as a field has been an important, if difficult to trace, influence on the development of PAD. In the seminal book *The Sciences of the Artificial*, Herbert Simon (1969, p. 113) argued for a science of design that is "tough, analytic, partly formalizable, partly empirical, teachable doctrine." The response to this call was design science. The purpose of design science is to produce and communicate evidence-based design propositions (Johannesson & Perjons, 2014) of the form, "If you want to achieve Y in situation Z, then apply intervention X." There are four defining characteristics of design science (Johannesson & Perjons, 2014) that influenced PAD. First, the goal of design science is to generate and test hypotheses about artifacts, such as assessment design patterns, as solutions to design challenges. Next, these hypotheses are tested and theory is built using rigorous research methods typically borrowed from the social sciences, including protocol analysis and field trials. Third, explicit prescriptions are offered on how to design and develop an intervention, such as game-based or three-dimensional science assessment. Finally, new findings on effective design are communicated to both practitioners (e.g., psychometricians) and researchers (e.g., learning scientists in areas such as learning progressions in science education).

This way of thinking in terms of design propositions supporting both the construction of assessments eliciting targeted knowledge and skills and the creation of evidentiary arguments supporting score interpretation is evident in PAD through the mindset of reasoning from imperfect evidence. In PAD, for example, design propositions undergird design patterns, and the latter are a link in the evidentiary argument supporting reasoning from imperfect evidence. These four defining characteristics of design science come into play not only in the design process, but also in the ongoing validation process, particularly with regard to the theory of action (addressed in a later section of this chapter).

PAD: ESSENTIAL ELEMENTS

To decide to engage in PAD does not mean that one has committed to a rigid set of processes in a particular order; PAD is not like putting together a piece of furniture from a kit with step-by-step directions. Rather, at its essence, it is "a mindset that results in a useful set of documentation" (M. Pearlman, personal communication, 2003). As described in the previous paragraph, that mindset is reasoning from imperfect evidence. And reasoning requires ruthless interrogation of ones' assumptions, always coming back to the essential question, What is my rationale for this decision or action? This question can be asked of any action within PAD, for example, What research has informed this definition of the construct? Are these the knowledge and skills valued in

the discipline? Why did I choose to use this item type for this target of measurement? What evidence will this item yield to help me place the student along the underlying proficiency trajectory? Why do I believe that is so? And so on. At its heart, the PAD endeavor involves a team of interdisciplinary players who are committed to this type of thinking and discourse throughout the design and development process. When the answers to these questions are transformed into reference documents and design tools, and then the resulting documentation is used as the infrastructure for the inferential and validation arguments that are articulated and refined continuously throughout the design and development process, then one has engaged in PAD.

In what follows, we have distilled PAD to six essential elements organized into three general steps. It is critical to keep in mind that we have organized in this way for ease and clarity of description, but in practice the lines between the steps are blurred as the elements (or, to be more specific, the artifacts representing the elements) are refined through the iterative design and development process. These elements and steps represent the design and development process up to the first field-testing event. After data are collected, those data are used to inform revisions to the artifacts. This feedback loop of using data to inform refinement of the artifacts continues through the life of the assessment program.

The use of PAD will vary depending on the training, expertise, and philosophical orientation of the individuals comprising the assessment design team and the particular needs of the assessment program. For example, PAD will be implemented differently by a team who has extensive experience applying ECD or assessment engineering in a variety of large-scale assessment contexts versus by a team whose focus of applying PAD, ECD, or assessment engineering has been in a single domain (e.g., scientific inquiry in chemistry) and whose work has been designing rich, performance-based tasks for the classroom. The former may use terminology like "student model," "evidence model," "observable evidence," "task model," "difficulty drivers," and "grammar," whereas the latter may use terminology like "focal knowledge, skills, and abilities," "unpacking," and "design patterns." These differences in terminology are simply different access points to the same desired outcomes of PAD, which we have distilled here into the six essential elements. Similarly, engaging in PAD will look very different when one is at the very beginning of designing an assessment versus whether one is using PAD to make improvements to an assessment program that has been operational for years. We posit that as long as some form of one or more of these essential elements is used—and the mindset is present—PAD will have some of the positive impact intended.

The three general steps are: (a) analyze the domain, (b) model the domain, and (c) create design patterns. As one moves through the steps, greater specificity and more attention to the constraints of the assessment are required. For example, in the domain analysis, one may articulate at a very high level the prioritized assessment targets (e.g., in mathematics, modeling, reasoning, and problem-solving are prioritized over procedural knowledge; algebraic reasoning is prioritized over geometric proofs). Then, as one moves into domain modeling, one must be much more specific to articulate the PLDs that will undergird the test and item specifications and support the intended score

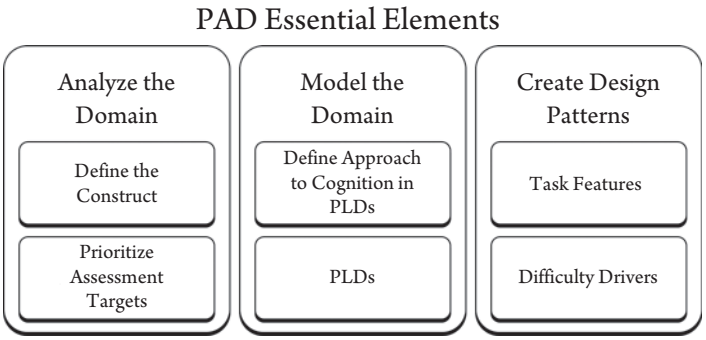

FIGURE 7.2

Principled Assessment Design Essential Elements

Note. PAD = Principled Assessment Design; PLD = Performance Level Descriptor.

inferences. Greater specificity still is required when creating design patterns because the constraints of the assessment must be considered when determining what kinds of tasks are best suited for the intended targets of measurement and how many tasks will be required to appropriately sample the domain and collect enough evidence to reliably support the intended claims about what students know and can do. Assessment design experts could reasonably disagree on when the assessment constraints should be considered in the design process. We argue that when an assessment is designed with the intention of cohering with instruction, the constraints of the assessment should be introduced at the domain model rather than during the domain analysis, and then the constraints are used more directly to shape the design pattern elements. In this way, the full breadth of the construct can be articulated and prioritized free of considerations of available item types and scoring mechanisms since those same constraints are not present in the classroom or in the discipline. Examples of richly articulated domain analyses can be found in the 2026 NAEP Mathematics and Reading Frameworks (NAGB, 2021a, 2021b), where the reader can readily envision reading and mathematics as sociocognitive and sociocultural endeavors in the classroom and in the real world without having to consider the respective constructs through a restrictive lens of constraints and features that are specific to the NAEP examination context.

See Figure 7.2 for the steps and elements of each step. Each will be discussed in detail in the following paragraphs.

ANALYZE THE DOMAIN

Define the Construct

It is unusual in the education context for the assessment developer to be charged with defining the construct, but when this is the case, the construct should be defined based on what we know about how students learn and build knowledge in the domain of interest and how that knowledge is used in the relevant discipline(s). In other words, for assessments of, say, biology, the construct should be defined based on what we know

about how students learn and build deep conceptual understanding in biology, as well as how this knowledge is used and employed in the biological sciences through the scientific practices. The role of learning science in the domain analysis cannot be understated for assessments that are intended to support inferences about what students know and can do, and even more so for assessments that are design to support instructional decisions (NRC, 2001; Perie & Huff, 2016). The theory of learning is the first segment of the assessment triangle from Pellegrino et al. (2011), cognition: What are our research-based hypotheses about how students learn and grow in the domain? It is this theory of learning that undergirds how cognitive complexity plays out in the domain model and, ultimately, the design patterns. For example, the NGSS are considered unprecedented in their complexity because they attempt to lay the groundwork for how science knowledge is not only acquired but also used in practice, which is why the standards are presented as performance expectations that reflect the integration of disciplinary core ideas (i.e., the learning that is required as the foundation for more learning), science and engineering practices (i.e., how knowledge is used and applied in the discipline), and crosscutting concepts (i.e., making connections across science domains is a way that students can make meaning of new information).

Another example of the essential role that learning science plays—or should play—in what we teach, learn, and assess has played out quite publicly between the mid-2010s and the mid-2020s in the United States (Goldstein, 2020; Hanford, 2018, 2019). As fourth-grade reading proficiency rates on NAEP continued to hover in the 30%–40% range for decades, Hanford (2018, para. 3) asked,

> How do we know that a big part of the problem is how children are being taught? Because reading researchers have done studies in classrooms and clinics, and they've shown over and over that virtually all kids can learn to read—if they're taught with approaches that use what scientists have discovered about how the brain does the work of reading. But many teachers don't know this science.

Although a number of empirical studies have shown that children need explicit phonics instruction to learn to read, "there is a history of ignorance, complacency and resistance in colleges of education with regard to disseminating this critical information to pre-service teachers" (Hurford et al., 2016, abstract). This same kind of ignorance, complacency, and resistance is evident in conventional approaches to assessment when it comes to building on learning science research in our design endeavors. As this example demonstrates, it is critical that scientific research about how students learn in the domain of interest be the foundation of our work in the domain analysis, which paves the way for modeling the domain, where we articulate with more specificity the approach to cognition in the performance levels as well as how students move from novice to mastery with regard to each grade-level expectation.

Similarly, when defining the construct, we must consider that the assessment should yield valid interpretations for all students: students from various sociocultural backgrounds, multilanguage learners, and students with disabilities. Understanding from the

outset what the research indicates about how students from these various populations relate to learning the domain of interest is critical to defining the construct in ways that do not automatically include construct-irrelevant features for these students (Hong & Lissitz, 2017; Randall, 2021; Solano-Flores, 2019; see also Ercikan & Solano-Flores, this volume).

Assessment Targets

Once the construct is defined in this way, then the prioritized assessment targets can be defined at a high level. As mentioned in the section "PAD: Essential Elements," we recommend that assessment constraints are not considered at this point, so that the assessment targets can serve not only as a basis for interim and summative assessment, but also for a variety of classroom and formative assessments. For example, in the redesign of the AP science courses, an approach was used to organize the domain into big ideas, enduring understandings for each big idea, and supporting understandings for each enduring understanding (Ewing et al., 2010). It was decided early in the design process that the target of learning, and therefore assessment targets, should be enduring understandings because these are the foundations that best prepare the learner for additional learning. Similarly, when the CCSS Mathematics were released, the guidance was for the major work of the grade, supporting standards, and additional standards to be proportioned 70%, 20%, and 10%, in curricular design, instruction design, and assessment design. These types of high-level guidance that are determined free of consideration of the interim or summative assessment constraints are critical to support coherence across instruction and assessment.

Most often in assessment of educational achievement, however, the analysis of the domain is provided to the test makers in the form of learning standards and test specifications that take the constraints and design features of the assessment into account. The constraints of the assessment program delimit the design features. Constraints and design features are determined by the answers to questions such as, but not limited to, the following.

The following are examples of constraints:

- How much time is there for design and development? What resources are available (e.g., personnel, funds)?
- What are the time limits for administration of the assessment? What is the expected turnaround time for score reporting?
- Must the assessment maintain a scale and/or comparability with an existing assessment, or is there freedom to depart?
- What item types are possible given the item authoring platform that will be used?
- What scoring mechanisms are available? For example, how much time and money exist for human raters, or must everything be machine scorable? What format(s) do the scoring mechanisms support (e.g., text, spoken word)?

The following are examples of design features:

- Is the assessment paper based or computer based?
- If computer based, is the assessment linear or adaptive?
- Does the client require a specified balance of item types?

We will return to the notion of constraints and design features in the section on "Design Patterns."

When the assessment designer is provided the learning standards (either with or without test specifications) as a starting place, the PAD process can start with modeling the domain. If the learning standards were not developed from a research base of how students learn, then the relevant learning science research must be integrated in the domain model and design pattern elements. Sometimes this process is called *unpacking* (Harris et al., 2016, 2019).

When test specifications are also provided as a starting place, excavating the implicit judgments that undergirded the myriad decisions that were made to get from learning standards to test specifications is of utmost importance in PAD. Otherwise, the assumptions about the relationships among the learning standards and the evidence yielded by an assessment with the said specifications are left unarticulated, which undermines the coherence required for a strong inferential argument and a strong validity argument to support the resulting interpretations about what students know and can do.

Model the Domain

As stated previously, various instantiations of PAD throughout the years have taken different perspectives on what activities are associated with each step of the process and what artifacts are created in that step. For example, Mislevy and Riconscente (2005) used PAD as a tool for developing rich, computer-based performance tasks of scientific inquiry. In their approach, the domain model includes, among other elements, the assessment argument in Toulmin form (1958, 2006) and the design patterns. However, for the purposes of this chapter and in the context of designing interim and summative assessments of educational achievement for which one of the primary purposes is to classify students categorically along a latent performance trajectory, the two critical elements that must be articulated after the domain analysis is complete and before task design can begin are: (a) an articulation of the approach to cognition and (b) development of the PLDs.

Approach to Cognition

It is not yet common practice for educational assessment programs to identify explicitly the model of cognition that is implicit in the PLDs, the tasks, the evaluation of student work, the scoring model, and, ultimately, the interpretation of what students know and can do (Ferrara & DeMauro, 2006; Huff et al., 2017). This is true despite the great advances that cognitive research in how students process information has made since the publication of the fourth edition of *Educational Measurement* in 2006 and that of

the current volume in 2025. When we pay more attention to the role of cognition in learning, we unleash enormous potential to improve assessment design and sharpen our interpretations about what students know and can do when we pay more attention to the role of cognition (Huff et al., 2010; Leighton & Gierl, 2011; Mislevy, 2006; Nichols & Huff, 2017; NRC, 2001; Snow & Lohman, 1989). If the role of cognition is not explicitly engineered into the assessment design process, then we risk the emergence of an implicit, fragmented, and flawed proxy "model," which most often leads to defaulting to the measurement of declarative or procedural knowledge or happenstance complexity (Huff et al., 2017).

We posit that given the constraints within which most educational assessments are designed and developed, it is likely to be a bridge too far to expect that a model of cognition be researched and defined a priori. However, it is essential that what we are calling the *approach to cognition* be articulated prior to (or in conjunction with) development of the PLDs. If not, then the assumptions about cognition will remain implicit in the individual minds of each player in the design endeavor, and it is unlikely that these implicit assumptions cohere. When models of cognition remain implicit in the minds of the various designers of PAD elements and consumers (primarily, educators), coherence is not achieved and the meaningfulness of the assessment results is compromised.

Before we describe what is meant by *approach* to cognition, a few comments on terminology are warranted. Table 7.1 gives a few illustrative examples of how related terms—*cognitive model*, *cognitive demand*, *cognitive complexity*, etc.—are defined by various authors. We recommend that for the purposes of most interim and summative

Table 7.1 Various Characterizations of Cognition Related to Assessment

Term	Citation	Definition
Cognitive model	Leighton & Gierl, 2007	A simplified description of human problem-solving of standardized educational tasks, which helps to characterize the knowledge and skills students at different stages of learning have acquired and to facilitate the explanation and prediction of students' performance
Theory of cognition	Huff et al., 2017	Prioritized value of knowledge and skills within a domain, organized to reflect structures and progressions along a learning trajectory, that articulates the precise knowledge and skills required to respond correctly to an assessment item
Cognitive demand	Perie & Huff, 2016	The degree to which tasks require more complex knowledge and skills for students to respond correctly and comprehensively
Cognitive complexity	Ferrara et al., 2014; Koedinger et al., 2012	The condition encoding requirements and the response requirements of the task; how many response operations are needed; interactions among different sources of complexity
Intrinsic cognitive load	Gillmor et al., 2015	Construct-relevant item features that contribute to item difficulty

educational assessments, becoming facile with these terms and definitions would be helpful but is not a requirement to achieve the baseline clarity and precision required for coherence. What is critical is that there is consensus on the assumptions undergirding student cognition for the domain of interest, that these assumptions are based in scientific research, and that the assumptions be explicitly reflected in the PLDs and design patterns. This degree of thought, discussion, and consensus among the design team regarding articulation of the approach to cognition reaps dividends as the team starts to author and review items from the perspective of whether the item yields sufficient evidence for the intended target of measurement, item clarity and quality, accessibility, sensitivity, bias, and cultural and linguistic responsiveness. When the PLDs and design patterns are authored with clarity on how student cognition is represented in the targets of measurement, conversations about construct-relevant versus construct-irrelevant variance start from a place of shared understanding rather than implicit assumptions.

There are at least three essential issues for the PAD team to debate and on which to reach consensus when building a shared understanding of the approach to cognition. The first is related to the fact that students can use many different cognitive processes and strategies to solve problems. For the intended purpose and use of this assessment, and given the prioritized knowledge and skills of the domain and the assessment targets that were defined in the domain analysis, what process and/or strategies will be assessed, if any? Or is it out of scope for the intended purpose and use of the assessment to prioritize process and/or strategies as targets of measurement? The intended purpose and use of the assessment will determine whether the process or strategy used to respond correctly to an item is also a measurement target. For example, in multiplication of multidigit integers, students may employ one of many strategies to correctly solve 125×10. One common strategy is to regroup, for example, $(100 \times 10) + (25 \times 10)$. Another common strategy is to factor, for example, $125 \times 2 \times 5$. Whether strategies are a target of measurement must be articulated a priori because this decision will have an impact on the cognitive approach, PLDs, and design patterns. It is worth noting that if it is determined that processes and strategies are not prioritized as targets of measurement, assessment designers will still need to articulate and discuss the processes and strategies used by students because they are directly related to preconceptions and misconceptions, which need to be articulated as part of the design patterns.

The second issue for the team to discuss is the verbs that will be used to represent skills and how those verbs are defined in terms of observable evidence. The importance of using clear explicit action verbs cannot be understated (Egan et al., 2012; Perie, 2006; Perie & Huff, 2016). For example, verbs such as "know" or "understand" are ambiguous and are not nearly as readily interpretable as "identify," "describe," and "explain." Another layer of clarifying verbs is to ensure that any synonyms are identified, for example, do we mean the same thing when we say *identify* versus *determine*? *Interpret* versus *analyze*? Perhaps one of the most critical elements of defining the approach to cognition is to define each verb in terms of observable evidence. In other words, if I asked a student to *interpret* a graph, what would I be looking for in the resulting interpretation

as evidence that the student interpreted rather than just, say, *described* the graph? What would I circle, underline, or highlight as evidence of an *interpretation*?

It is crucial that this thought process and discussion among the design team members occur without consideration of the assessment constraints, for two reasons. First, the approach to cognition is the first essential element of the domain model as it will directly inform PLDs. And, as cautioned earlier about introducing assessment constraints too soon, if the PLDs are to serve both curricular design and assessment design, and have reasonable generalizability to the domain of interest, they need to be written with applicability in the classroom as well as for the assessment (which does need to be addressed during PLD development, but not during approach to cognition; see the next section).

Second, suppose the assessment must be limited to two item types: multiple choice and short fill in the blank. If the team discussion about observable evidence of *interpretation* assumes the constraints of a particular item type, then the discussion is actually about what can be done within those item-type constraints. Establishing the key observable elements of the skill *interpret* should lead to the design of the item type or the selection of the best item type available to elicit the observable evidence of the skill. An example of this will be provided in the section on "Design Patterns."

The third issue for the design team to discuss is: What assumptions do we have about how knowledge and skills relate to each other and cognitive demand? For example, in the assessment of historical thinking (as opposed to the assessment of recalling historical facts), it is generally agreed that the historical thinking skill "*identify and explain* historical developments" has less cognitive demand than "*analyze* patterns and connections between and among historical developments" regardless of the historical content, for example, pre-Columbian migration patterns or Victorian morality (Ercikan & Seixas, 2015), whereas in the sciences, the pairing of skill and content matters (Ewing et al., 2010). For example, the cognitive demand of accurately *describing* Heisenberg's uncertainty principle is likely higher than the cognitive demand of, say, *analyzing* the equation force = mass × acceleration, even though one would normally assume that *analyzing* requires more cognitive demand than *describing*.

Using skill taxonomies such as Webb's Depth of Knowledge (DOK), Bloom's taxonomy, SOLO taxonomy (Anderson & Krathwohl, 2001; Biggs & Collis, 1982; Webb, 2005), or others can be useful in considering the cognitive approach but are not sufficient, because the cognitive approach must be defined in light of the domain analysis—that is, in light of the prioritized knowledge and skills and in response to how students learn in the domain of interest. Given that most assessments of educational achievement are based on a set of learning standards that are not developed using PAD, it is likely that the design team will need to create an approach to cognition that is complementary to the learning standards. For example, suppose that the learning standards use a preponderance of ambiguous or less challenging verbs to represent skills (e.g., know, understand, identify, determine). The design team will have an opportunity to probe for clarity and desired level of rigor with stakeholders during the PLD development

process. Clarity will also be required when the learning standards include skills that are not generally fully assessed via typical exam constraints (e.g., evaluate, design, explore, create). The designers and stakeholders will need to reach agreement on the acceptable and reasonable evidence of these learning standards given the exam constraints.

Building consensus for—and documenting—the cognitive approach before developing the PLDs is critical because any latent ambiguity will cascade to the PLDs, the design patterns, and, ultimately, inferences about students. For assessment designers working on an assessment program that is already operational, where PAD was not present in the original design, all is not lost. There are at least two ways to infuse research-based approaches to cognition into an operational assessment program. First, most assessment programs need some new items each year. These consensus-building discussions that reveal latent assumptions about skills, verbs, and the relationships among skills and content can be used to improve item development specifications and item review criteria. Second, most educational assessment programs designed to classify students conduct standard setting on a specified cadence of every 3–5 years, or when needed, and most assessment programs take this as an opportunity to update PLDs (see Ferrara et al., this volume, for a discussion of standard setting). Updating PLDs is the perfect opportunity to infuse a transparent approach to cognition into the assessment design.

PLDs

PLDs are the claims about what students know and can do as they grow in mastery in the domain. In PAD, PLDs must be developed in advance of design patterns because they are essential elements in item design, scale development, and standard setting (Bejar, 2010; Bejar et al., 2007; Egan et al., 2012; Huff et al., 2017; Schneider et al., 2013). Generally speaking, PLDs reflect how knowledge and skills become more sophisticated as the achievement levels progress from representing *Novice* or *Emergent Mastery* through *Advanced Mastery*. The PLDs should emerge directly from the domain analysis and the approach to cognition that have been previously determined in the PAD process, both of which must emerge from the science of learning in the domain. Without this coherence, we risk that the inferences about where students are along the learning trajectory will not be useful to educators (Bejar, 2010; Frederiksen & Collins, 1998; Nichols et al., 2009; Schneider, 2017).

Egan et al. (2012) distinguished between the following PLD types and their uses in test development and score reporting: policy PLDs, range PLDs (RPLDs), target PLDs, and reporting PLDs. We discuss each in turn. More emphasis is given to the RPLDs given their role in item design, standard setting, and scale development.

Policy PLDs

Policy PLDs are high-level descriptions of what students should know and be able to do at each performance level. This description is developed for each domain and is typically common across grades. The policy-level definitions from NAEP Grade 4 Mathematics are shown in Figure 7.3. Policy-level PLDs are critical for setting the baseline expectations of the more detailed range, target, and reporting PLDs.

Grade 4

NAEP Basic **(214)**	**Fourth-grade students performing at the *NAEP Basic* level should show some evidence of understanding the mathematical concepts and procedures in the five NAEP content areas.** Fourth-graders performing at the *NAEP Basic* level should be able to estimate and use basic facts to perform simple computations with whole numbers, show some understanding of fractions and decimals, and solve some simple real-world problems in all NAEP content areas. Students at this level should be able to use–though not always accurately–four-function calculators, rulers, and geometric shapes. Their written responses will often be minimal and presented without supporting information.
NAEP Proficient **(249)**	**Fourth-grade students performing at the *NAEP Proficient* level should consistently apply integrated procedural knowledge and conceptual understanding to problem solving in the five NAEP content areas.** Fourth-graders performing at the *NAEP Proficient* level should be able to use whole numbers to estimate, compute, and determine whether results are reasonable. They should have a conceptual understanding of fractions and decimals; be able to solve real-world problems in all NAEP content areas; and use four-function calculators, rulers, and geometric shapes appropriately. Students performing at the *NAEP Proficient* level should employ problem-solving strategies such as identifying and using appropriate information. Their written solutions should be organized and presented both with supporting information and explanations of how they were achieved.
NAEP Advanced **(282)**	**Fourth-grade students performing at the *NAEP Advanced* level should apply integrated procedural knowledge and conceptual understanding to complex and nonroutine real-world problem solving in the five NAEP content areas.** Fourth-graders performing at the *NAEP Advanced* level should be able to solve complex and nonroutine real-world problems in all NAEP content areas. They should display mastery in the use of four-function calculators, rulers, and geometric shapes. The students are expected to draw logical conclusions and justify answers and solution processes by explaining why, as well as how, they were achieved. They should go beyond the obvious in their interpretations and be able to communicate their thoughts clearly and concisely.

FIGURE 7.3

National Assessment of Educational Progress Grade 4 Mathematics Policy-Level Definitions

Note. NAEP = National Assessment of Educational Progress.

RPLDs

RPLDs are detailed descriptions of what students should know and be able to do in each performance level. RPLDs acknowledge that students within a given performance level represent a range of performance; for example, students at the low end of *Proficiency* have a different level of knowledge and skill than students at the high-end range of that level. Several factors should be considered in the organization and format of the RPLDs. First, the definition of the construct and prioritized assessment targets (the two essential elements of the domain analysis) and the approach to cognition (the first essential element of the domain model) should inform the structure and grain size of the RPLDs. Both the Partnership for the Assessment of Readiness for College and Career (PARCC) and the New York State Testing Program (NYSTP) used different elements of PAD in their development (Huff et al., 2017). In each case, the testing programs were provided with the domain analysis in the form of the CCSS that were designed to inform both curriculum and assessment and contained some high-level guidance on which standards should take priority. The CCSS also provided a starting

PARCC PLD Example

Grade 3 Math: Content (Sub-Claim A) The student solves problems involving the Major Content for the grade/course with connections to the Standards for Mathematical Practice.			
Exceeds Expectations	**Meets Expectations**	**Approaches Expectations**	**Partially or Does Not yet Meet Expectations**
Products and Quotients: 3.OA.1, 3.OA.2, 3.OA.4, 3.OA.6, 3.OA.7-1, 3.OA. 7-2			
Understands and interprets products and quotients of whole numbers.	Interprets products and quotients of whole numbers.	Interprets products and quotients of whole numbers.	Determines products and quotients of whole numbers within 100.
Determines the unknown whole number in a multiplication or division problem by relating multiplication and division. Both factors are greater than 5 and less than or equal to 10.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division. One factor is greater than or equal to 5.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division, with both factors less than or equal to 5, or with one factor of 10.	Determines the unknown whole number in a multiplication or division problem by relating multiplication and division, with both factors less than or equal to 5, or with one factor of 10.
Represents a multiplication or division situation as an equation.			
Accurately multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	Accurately multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	Multiplies and divides within 100, using strategies relating multiplication and division or properties of operations.	
Multiplication and Division: 3.OA.3–1, 3.OA.3-2, 3.OA.3-3, 3.OA.3-4			
Uses multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Both factors are greater than 5 and less than or equal to 10.	Uses multiplication and division within 100 to solve word problems involving equal groups and arrays. One factor is greater than or equal to 5.	Given a visual aid, uses multiplication and division within 100 to solve word problems involving equal groups and arrays, with both factors less than or equal to 5, or with one factor of 10.	Given a visual aid, uses multiplication and division within 100 to solve word problems involving equal groups. Both factors are less than or equal to 5, with both factors less than or equal to 5, or with one factor of 10.

NYSTEP PLD Example

Grade 3 Mathematics Performance Level Descriptions				
Cluster	Performance Level 4	Performance Level 3	Performance Level 2	Performance Level 1
Students represent and solve problems involving multiplication and division. (3.OA.1-4)	Interpret and represent products and quotients of whole numbers. Determine the unknown whole number in a multiplication and division problem by relating multiplication and division. Represent a multiplication or division situation as an equation. Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Identify proper context given a numerical expression involving multiplication and division. Both factors are less than or equal to 10.	Interpret products and quotients of whole numbers. Determine the unknown whole number in a multiplication or division equation relating three whole numbers by relating multiplication and division. Factors are greater than 5 and less than 10. Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Both factors are less than or equal to 10.	Interpret products of whole numbers. Determine the unknown whole number in a multiplication equation by relating multiplication and division. Limit to factors less than or equal to 5. Given visual models and/or manipulatives, use multiplication and division within 100 to solve word problems involving equal groups and arrays. Both factors are less than or equal to 10.	Given visual models and/or manipulatives, interpret products of whole numbers with factors less than or equal to 5. Determine the product in a multiplication equation with whole number factors less than or equal to 5. Given visual models and/or manipulatives, compute products within 25 in the context of word problems.
Students understand properties of multiplication and the relationship betwen multiplication and division. (3.0A.5, 6)	Justify the use of properties of operations (commutative, associative, and distributive) as strategies to multiply. Restate a division problem as an unknown factor problem, and explain the relationship between division and finding an unknown factor.	Apply properties of operations (commutative, associative, and distributive) as strategies to multiply. Restate a division problem as an unknown factor problem.	Apply the commutative property as a strategy to multiply. Restate a division problem as an unknown factor problem.	Given visual models and/or manipulatives, identify equivalent expressions that illustrate the commutative property within 10.

FIGURE 7.4

Range Performance-Level Descriptor Examples

place for the approach to cognition in the form of the mathematical practices and an articulation of the relationship among the reading standards and text complexity. Figure 7.4 shows the different ways that the RPLDs were organized by PARCC and NYSTP; different approaches were used, but each was reasonable. For PARCC, the performance levels were organized by standard, whereas for NYSTP, cluster was the organizing unit. NYSTP also provided a bit more granularity in the PLDs than PARCC.

It is important to consider whether the RPLDs will be used to support curriculum, instruction, and broader assessment uses in the classroom or just the interim or summative assessment that will need to meet large-scale assessment psychometric and validation requirements. In a coherent educational system, the promise is that the RPLDs will be used to support both the classroom and the large-scale assessment, and if that promise is to be realized, then issues related to the verbs used to represent skills and the difference between assessment constraints and classroom constraints need to be considered in advance of RPLD development.

For example, take the two skills "identify" and "name" into consideration with regard to different kinds of parallelograms. Let's suppose identify is defined in the approach to cognition as a student's ability to select the correct term from a set of options (e.g., square, rectangle, rhombus) when presented with a shape with particular features, whereas name is defined as the student's ability to independently generate the correct term. Unless ground rules are laid in advance, the RPLD developers could introduce unnecessary limits on the RPLDs by taking the assessment constraints into account. That is, if it is known in advance that the assessment will not be able to accommodate constructed response or speech recognition for name, then some would argue that because the RPLDs will serve item design, standard setting, and scale development, then the skill "name" should not be used in the RPLDs. However, that would make the RPLDs less useful for the classroom, which does not have the same constraints as the assessment setting. As such, we recommend that when the intent is for the RPLDs to serve both the classroom and the large-scale assessment, the RPLDs be written without the assessment constraints taken into account. We argue that the resolutions required for the RPLDs to serve item development, standard setting, and scale development can happen at a later time. For standard setting, a skills glossary can be provided for panelists that indicates, for example, "for the purposes of this methodology, assume the verb *identify* each time the verb *name* is used," or the verbs can be changed accordingly in the PLDs that will be used in standard setting. For item and scale development, the design pattern is where resolutions between targets of measurement (skills, knowledge) and assessment constraints (e.g., item types) are addressed.

Claims about students in the RPLDs should reflect the approach to cognition previously articulated, reflect the knowledge or content prioritized in the domain analysis, and identify any relevant context that has an impact on the cognitive demand, and therefore the performance level in which the student claim is situated (Egan et al., 2012; Ferrara & Steedle, 2015; Ferrara et al., 2015; Schneider et al., 2013; Valencia et al., 2014). Before giving an example, it is important to note that the development of

the RPLDs will likely require some iteration with the approach to cognition. It is still advisable to document the approach to cognition prior to RPLD development so that all are working from a shared understanding and a shared vocabulary, which will facilitate productive iteration as opposed to unproductive churn.

Table 7.2 provides an example of RPLDs where the approach to cognition, knowledge (or content), and context is embedded in ways that are intended to model a progression of cognitive demand, from early grade-level mastery (left column) to advanced grade-level mastery (right column).

Typically, students in the earlier stages of mastery need reduced cognitive load to demonstrate what they know and can do with grade-level content and skills, so it is critical to include this kind of context at the appropriate RPLD performance level to differentiate on grade-level performance with scaffolds from below-grade-level work.

Target PLDs

Rather than the range of performance within a particular level, the target PLDs (sometimes referred to as the threshold PLDs) focus on the distinguishing characteristics

Table 7.2 Example Range Performance-Level Descriptors

Approach to Cognition: Verb Used to Represent the Skill		
Knowledge/content: The similarities and difference among similes, metaphors, allusions, alliteration		
Context: Text (stimuli) can require different levels of inference		
Early grade-level mastery	Mid-grade-level mastery	Late grade-level mastery
Student identifies an author's use of easily inferred figurative language or other literary devices (e.g., similes and metaphors) in a literary text.	Student identifies an author's user of figurative language or other literary devices (e.g., allusions and alliteration) and interprets how this language contributes to the meaning of a literary text (e.g., how a metaphor expresses a theme).	Student identifies an author's user of figurative language or other literary devices and evaluates how this language contributes to the meaning of a literary text that is complex and requires a high degree of inference.

at the "threshold" of each performance level. This facilitates necessary discussions in various standard-setting procedures where panelists need to reference a shared understanding of what knowledge and skills represent a student who is "just barely" in one level as opposed to the one just preceding it.

Reporting PLDs

Once the cut scores are finalized, any adjustments that need to be made for reporting what students in each performance level know and can do should be made. For example, a decision could be: Should the reporting PLDs reflect the minimally sufficient knowledge and skills of the placement level or the typical knowledge and skills of the placement level?

DESIGN PATTERNS

The third step in PAD—creating design patterns—has two essential elements: (a) an articulation of the task[3] features that are required to elicit the evidence necessary to support the intended claim about the student as articulated in the RPLDs and (b) an identification of the task features that impact cognitive demand (frequently referred to as difficulty drivers) and, by extension, item psychometric characteristics and scale properties. Both of these essential elements build on essential elements that have been articulated previously in the PAD process: the definition of the construct, the prioritized assessment targets, the approach to cognition, and the RPLDs. There are many other elements of the assessment design endeavor that will need to occur after RPLDs are authored before an assessment is ready for field testing; Figure 7.5 provides a summary.

We have chosen to highlight task features and difficulty drivers as essential elements because they are the unique features of PAD that typically remain implicit in conventional approaches to assessment design. As with other aspects of PAD, the

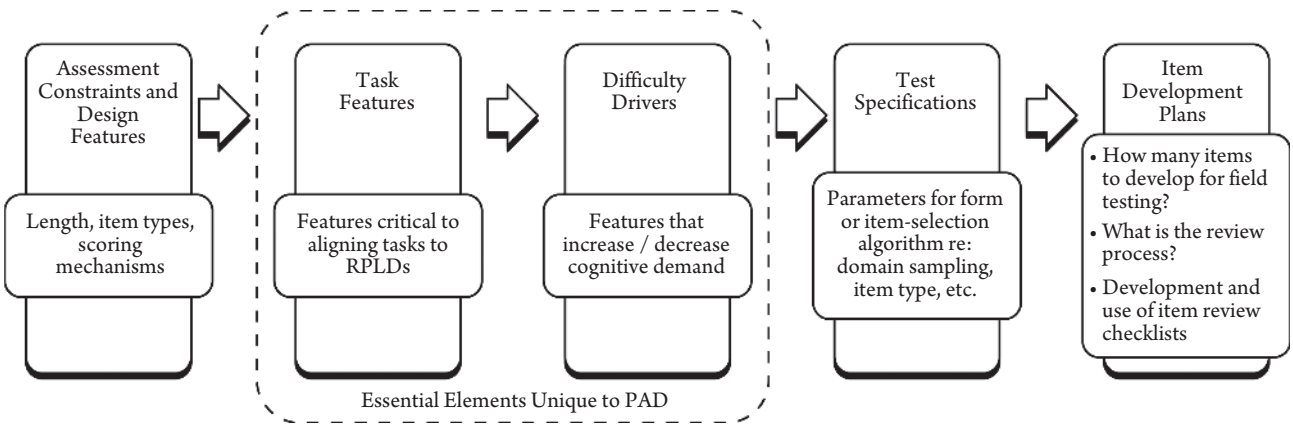

FIGURE 7.5
Remaining Steps in Assessment Design
Note. PAD = Principled Assessment Design; RPLD = range performance level descriptor.

terminology for the process of task design used in various publications and applications can be daunting and off-putting (e.g., design patterns; task models; task features; task specifications; variable features; fixed features; focal knowledge, skills, and abilities; assessment framework). In an attempt to clarify, we posit that some of the key benefits of PAD can be achieved if the task features and difficulty drivers are carefully articulated, used to develop items, and refined accordingly as data are analyzed as part of iterative, ongoing design. For operational testing programs not developed under PAD, infusing task features and difficulty drivers into the item design specifications can be done over time.

One of the most challenging aspects of creating design patterns is what grain size to start with: A single learning standard? An RPLD component? A single target of measurement for a single item? As Hendrickson et al. (2010) described, these decisions are arbitrary and the process is iterative. A good starting place will be influenced by the structure of the RPLDs and the assessment constraints and design features. For example, the NYSTP Grade 3 Mathematics RPLDs are organized by cluster (an organizing feature within CCSS) and then become more granular as the progression is articulated against four performance levels (see Figure 7.4). The first row of the first cluster has to do with *products and quotients of whole numbers* and is ordered from below grade-level performance to advanced grade-level performance as such:

- Level 4: Interpret and represent products and quotients of whole numbers.
- Level 3: Interpret products and quotients of whole numbers.

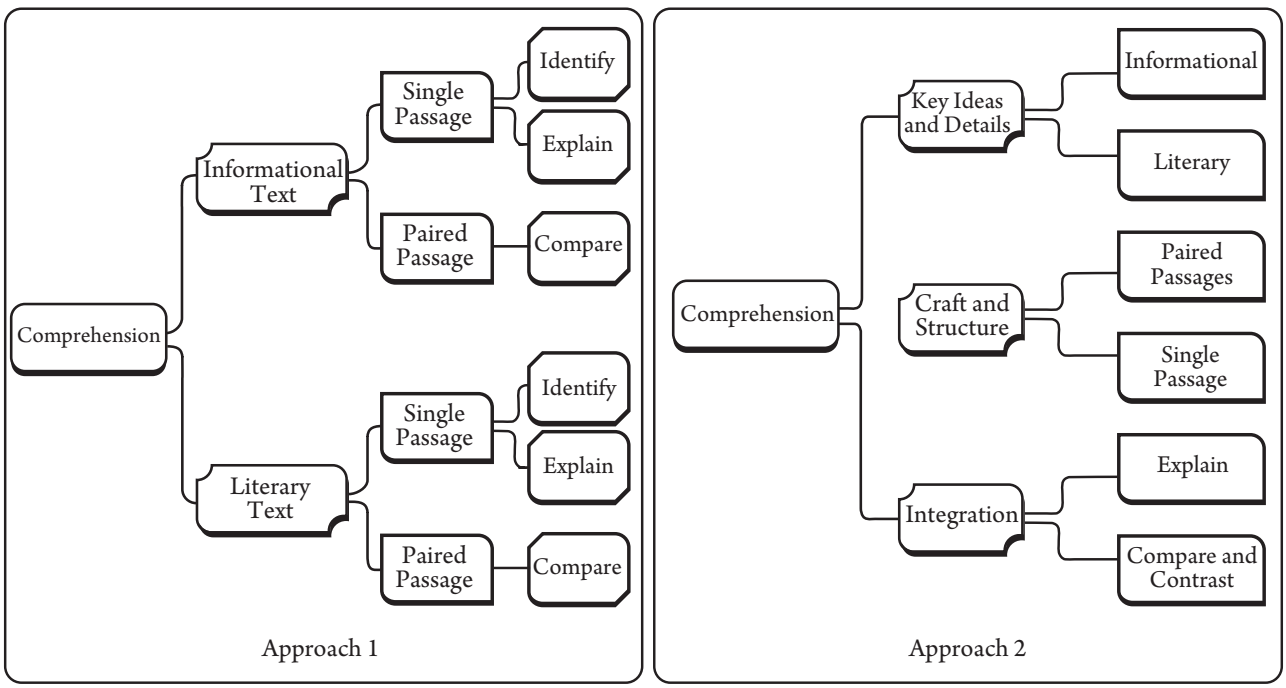

FIGURE 7.6
Design Pattern Organization and Grain Size

- Level 2: Interpret products of whole numbers.
- Level 1: Given visual models and/or manipulatives, interpret products of whole numbers with factors less than or equal to 5.

It seems that the first row of the first cluster would be a likely place to start and yield a useful grain size for a design pattern. However, if time and resource constraints are a factor, a design pattern that applies to the full cluster will be sufficient and better than no design pattern at all. Figure 7.6 provides an example of two different approaches to creating design patterns for a reading comprehension domain. In the first approach, the design patterns are organized by text genre, number of passages in the stimuli, and then specific skill. In the second approach, design patterns are organized by skill clusters and then by other features depending on the cluster. The take-home message from this example is twofold: There is variability and choice in the grain size of design patterns depending on the specific needs of the testing program and the resources available; and both the organization and the grain size of the design patterns are arbitrary. What is essential is that the task features that anchor items to the RPLDs and the difficulty drivers that impact cognitive demand are articulated prior to beginning item development. Although task features and difficulty drivers may change depending on whether one is working with a single informational passage where the target of measurement is identify versus explain, if there are not resources to create design patterns at that level of specificity, then having a design pattern one or two layers up is still beneficial. We must start where we can and claim small PAD victories along the way.

Task Features

Articulation of the task features provides the essential, explicit link between the RPLDs and the tasks to be written by explicating the features of stimuli (e.g., text, graph, video, map, diagram), prompt (the question or problem to be solved), key (the correct answer), distractors (incorrect answers based on preconceptions and misconceptions), and other ways in which unique aspects of the item type can be manipulated to help ensure that each item captures evidence relevant to placing the student along the latent performance continuum (Hendrickson et al., 2010; Luecht, 2013, 2019). Each feature of the task contributes to the cognitive demand of the task (Ferrara et al., 2015; Koedinger et al.; 2012; Schneider & Johnson, 2019), so each must be considered carefully in relation to the RPLDs. The constraints of the testing program will delimit each of the task features. For example, an assessment program that is restricted to machine-only real-time scoring and must be relatively short in duration (e.g., an adaptive interim assessment) will have task features that differ from an assessment program—of the same content and skills—that includes, say, essays or computer-based simulations. Each decision about task features should be warranted by the learning science research that undergirds the domain analysis and the domain model.

For example, a design pattern for the RPLDs shown in Table 7.2 would need to address, at minimum, the task features that differentiate performance between the levels with greater specificity. Currently, the RPLDs suggest that similes and metaphors

are associated with a different RPLD than allusions and alliterations. Given that similes, metaphors, allusions, and alliterations are provided as examples, what about other figurative language devices such as personification or hyperbole? Is each on grade level for this standard, and if so, are they associated with a particular RPLD? Other features that should be addressed in the design pattern include the features of text that must be present to support the kinds of questions that students will be asked to respond to in relation to the text, such as, What are features of text that make figurative language "easily inferred" as opposed to requiring a "high degree of inference"?

Also to be addressed in the design pattern are how different item types that are available for the assessment relate to the intended targets of measurement and performance levels. For example, the assessment design team may determine that the skill identify can be assessed with fidelity with either multiple choice or drag and drop, whereas the skill evaluate requires students to highlight relevant text in the passage. In each case, the rationale for the use of the item type(s) for the skill should be documented in the design pattern, as well as any specific features of the item type that must be present to elicit evidence for that particular skill. In addition, we need to refer to the approach to cognition when creating design patterns. In the approach to cognition, we define the observable features of the skills, in this case what constitutes an *identification* and an *evaluation*. One common characteristic of evaluation that distinguishes it from identification is a judgment of quality. In the design pattern, we articulate how the task features elicit from the student evidence of the RPLD claim (or part of the claim) that is the target of measurement, in this case, evidence that gives one confidence that the student is *evaluating* rather than *identifying*.

Difficulty Drivers

In the previous section, we asserted that an essential element of design patterns is to articulate the task features that elicit evidence that a student is in one placement level versus another. When the RPLDs are constructed carefully, the distance to get from RPLDs to design patterns is a few steps rather than a journey. A second essential element of design patterns is difficulty drivers, that is, What are the task features that we can manipulate to alter cognitive demand but stay within the intended RPLD? Recall that any performance level represents a range of proficiency, and items are needed that represent the range. Identifying the task features that can be manipulated to adjust cognitive demand is helpful to support development of items that will span a range of observed difficulty. When considering the relationships among task features and cognitive demand, it is imperative to consider the diversity of the student population such that we are avoiding construct-irrelevant difficulty for students from various sociocultural backgrounds, multilanguage learners, or students with disabilities.

Figure 7.7 shows an excerpt from a detailed design pattern from an AP Biology research study on automatic item generation (Huff et al., 2013). Here, the researchers identified six ways to manipulate the difficulty of items while still targeting the claim within the *Basic* performance level. Also noteworthy is that the researchers identified two task features that do not impact cognitive demand, which is also helpful information. Note that evidence

Construct Identifier:	Biology
Primary Context:	Cell division
Competency Claim	The student can construct explanations of how DNA is transmitted to the next generation via the processes of mitosis, meiosis, and fertilization.
Proficiency Level	Basic
Evidence Documentation	
1.	Description of the purpose of mitosis and meiosis
2.	Description of the products of mitosis and meiosis
3.	Description of the behavior of the chromosomes during the phases of mitosis and meiosis
4.	Explanation of the processes of mitosis and meiosis
5.	Comparison and contrast of the processes of mitosis and meiosis
6.	Use and recognition of vocabulary specific to cell division
Manipulable features of complexity	
1.	Type of cell division (mitosis is simpler than meiosis)
2.	Number of steps in process (mitosis has fewer steps than meiosis)
3.	Type of statement/alternative (definition is less challenging than explanation)
4.	Use of vocabulary particular to cell division will increase complexity: ploidy, tetrads, synapsis, crossing over, sister chromatids, homologous chromosomes, segregations, equatorial plate, cytokinesis
5.	Phase of cell division in question; the events in some phases are more conceptually difficult than the events of other phases
6.	Making a comparison (more challenging) vs. selecting a true statement (less challenging)
Features irrelevant to complexity	
1.	Number of chromosomes in a cell
2.	Type of organism in which the processes occur

FIGURE 7.7
Snapshot of Design Pattern With Manipulable Features of Complexity (Difficulty Drivers)

documentation in Figure 7.7 is the observable evidence of an explanation within this particular context of cell division; this is another example of how different approaches to PAD have the essential elements associated with different processes. No matter the differences, the intent is the same: to make transparent all that is typically implicit in task design so that we can debate our assumptions, document our consensus, and use those documents as design tools and as evidence of our inferential chain of reasoning.

Ferrara et al. (2014) worked with a team of researchers across many testing organizations and practitioners across many states to develop an approach to considering cognitive complexity for the CCSS. The goals for these new metrics were to provide a systematic and replicable way of determining cognitive complexity for each task and to provide measurement precision at all levels of the score scale. Figures 7.8 and 7.9 provide taxonomies of cognitive complexity for English language arts (ELA)/literacy and math, respectively. Note that for ELA/literacy, the complexity and length of the text is considered a source of cognitive complexity along with the processing demands of the task (e.g., the

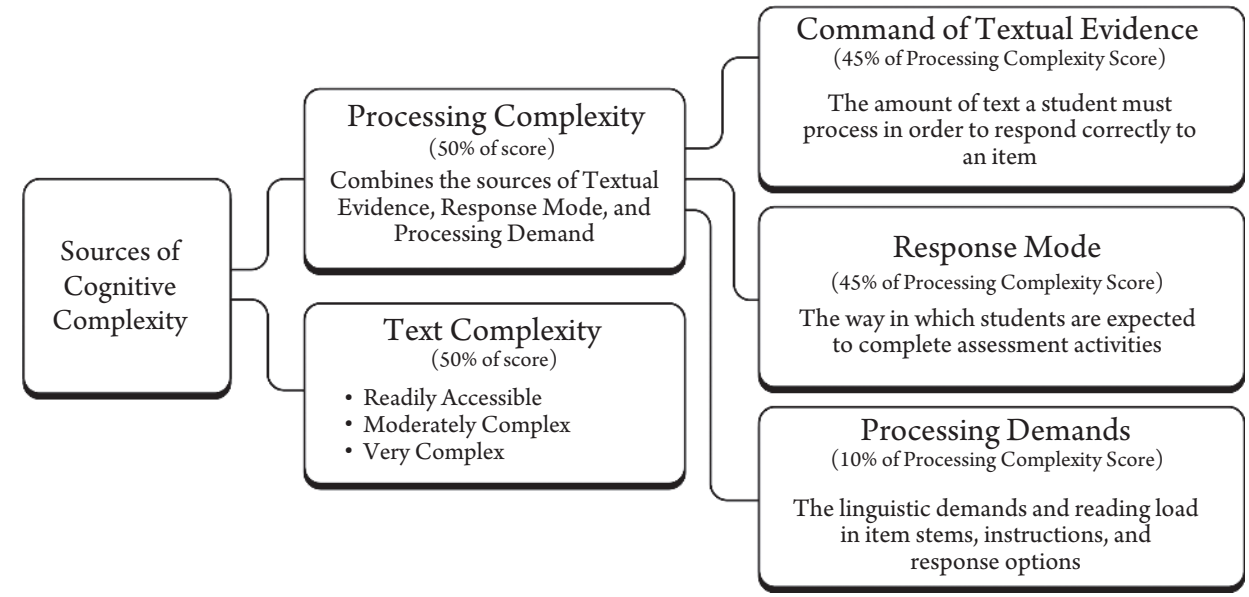

FIGURE 7.8
Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers English Language Arts/Literacy

stimulus and prompt presented to the student), as well as the response demands of the student (e.g., selected response versus constructed response). Similarly, for math, both stimuli demands and response demands are considered sources of complexity, as well as which mathematical practice is required in the task. For both ELA and math, there exist research-based rules for how to roll up these various features into a single metric of low, moderate, and high cognitive complexity that can be used in design patterns as a more sophisticated approach to difficulty drivers than conventional practice.

While this section has focused on design patterns, PAD does not negate the need to implement training to produce items of good technical quality and fairness. Readers desiring additional information in these areas should refer to Schmeiser and Welch (2006) or the *Handbook of Test Development* (Lane et al., 2016). The development of items must include focus on the accuracy, clarity, and accessibility of the items; the items being free of sensitivity and bias; and the items being culturally and linguistically responsive. We argue that when the PAD mindset of reasoning from imperfect evidence is shared by the design team and some of the PAD essential elements are developed and used as design tools, the resulting items have a strong pedigree—that is, a strong evidential argument regarding the validity of inferences we need to make about what students know and can do.

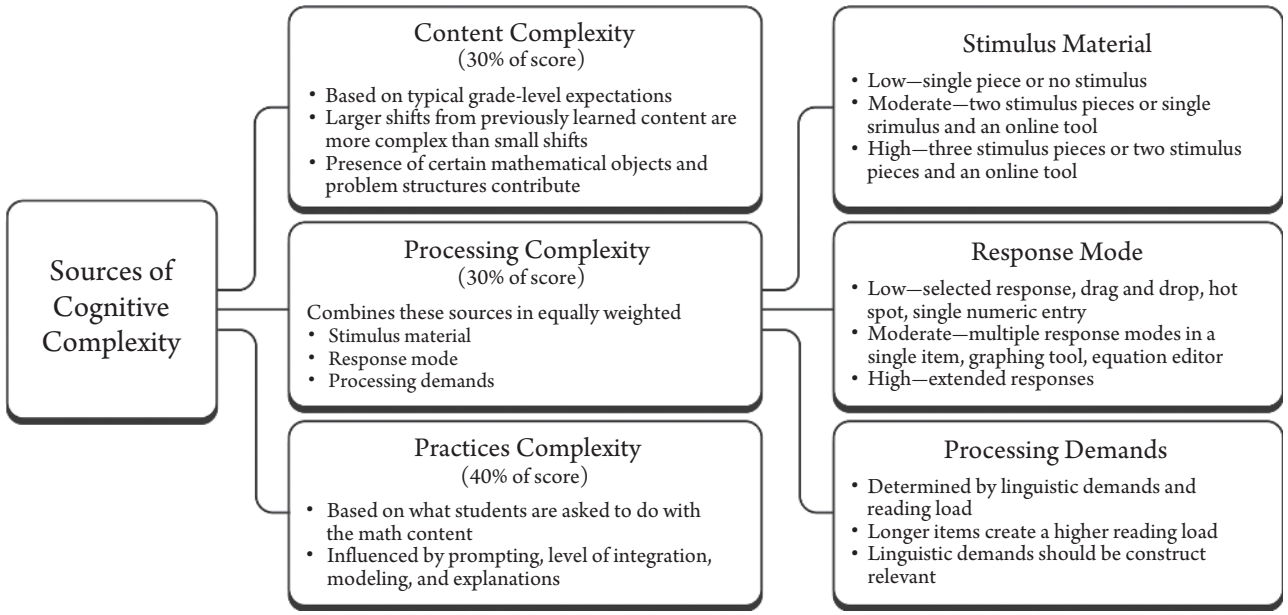

FIGURE 7.9
Sources of Cognitive Complexity for Partnership for Assessment of Readiness for College and Careers Math

INTEGRATION OF DESIGN AND PSYCHOMETRICS

One of the primary advantages to using PAD is the integration of several aspects of test design that are distinct and even siloed in conventional approaches. Integration of skills and content into claims about students, integration of RPLDs into item design, and, finally, integration of item design and scale design are beneficial in multiple ways and represent a true evolution in the assessment industry.

Although standards-based assessments or, more specifically, assessments intended to support inferences regarding the degree to which students have met learning objectives have grown in demand and prominence from the 1980s to 2025, the publication date of this volume, the way we design and develop large-scale educational assessments is still largely situated within philosophies, structures, and processes that evolved to support assessments for which normative information was the primary inference of interest, such as percentile rank with regard to a specific population. Item development processes and psychometric research on these norm-referenced tests—SAT, ACT, Iowa Tests (formerly Iowa Tests of Basic Skills), TerraNova, SAT-10, and GRE, to name a few from the field of education—have informed much if not most of educational

measurement and educational test development. When testing companies—and the psychometricians and item writers[4] in those companies—shifted from building "off the shelf"[5] or selection tests to tests based on state-specific learning standards with the primary purpose of classifying students into proficiency categories, the required shifts in philosophy and practice were minimal at best.

The primary metrics of conventional norm-referenced K–12 tests to support inferences about students are percentile-rank based on a nationally representative population and grade-equivalent scores—another norm-based metric devoid of any connection to learning standards—not inferences about what students know and can do with regard to a set of learning standards that represent what is valued in the classroom and the discipline. For college or graduate school admission tests, the primary metric to support inferences is, again, percentile-rank based on a nationally representative population, and the primary validity evidence is prediction of future grades. From a business perspective, it only made sense to keep as many practices and processes in place as possible during this pivot into the new space of state-specific standards-based testing.

Many test designers and psychometricians, with the best intentions, likely assumed that the only necessary changes to traditional practices concerned producing what the states needed to meet federal requirements: a post hoc alignment of banked items from these off-the-shelf tests to the state's learning standards and a standard setting after the first operational administration to establish cut scores for the performance levels. As discussed in the second paragraph of the chapter, the cry for a different philosophy and practice for educational assessment design gained strength and numbers during this era and culminated with the NRC's *Knowing What Students Know* (NRC, 2001). At the time of this writing over 20 years later, practitioners are either still fighting for evolution in the direction of PAD for assessments designed to tell us about what students know and can do or maintaining status quo either out of a lack of knowledge of a different way or out of a belief that the cost of evolution in design practices is neither warranted nor worthwhile from a business perspective. The authors of this chapter are firmly in the first category.

As mentioned at the beginning of this chapter, the conventional approach to test development, if viewed from a high level, may not appear to differ from a PAD approach. For example, below is a synopsis of the main phases of the test development process as outlined by Schmeiser and Welch (2006):

Step 1: Identify intended purpose and use

Step 2: Determine test design features and constraints (e.g., time to test, time to score, paper delivery or computer based, item type balance)

Step 3: Test specifications

Step 4: Item development

Step 5: Field testing

Step 6: Evaluation of test results

The PAD approach is essentially the same as the conventional approach when viewed at this grain size, especially for Steps 1, 5, and 6. When one goes into the details of the processes and documentation of Steps 2–4, however, there are many differences, especially for educational achievement tests that require performance levels and standard setting. One key operational and pragmatic differentiator between conventional approaches and PAD that has a ripple effect throughout the design and development process—as well as implications for the inferential and validity arguments—is that the former has distinction, if not fragmentation, in areas whereas the latter has integration and coherence. Three related areas of integration that are hallmarks of the PAD approach are: (a) the integration of content and skills to be assessed in the test specifications; (b) the integration of PLDs into the item design process; and (c) using PAD to engineer desired scale properties. These types of integration require a deeper partnership among item writers and psychometricians—as well as representatives from other related disciplines who should also have a seat at the table—during the design phase of the assessment than is typically the case.

Data to Inform Design and Development

In Schmeiser and Welch (2006), field testing is noted as Step 5. Field testing (sometimes referred to as field trials, pretesting, or item tryouts) is generally defined as when items are administered to a sufficient number of students to support initial psychometric analyses. New or trial items do not count toward a student's score, primarily to avoid miscalculating the student's score should subsequent psychometric analyses reveal a flaw with the item or because the trial item is not yet properly scaled with all the other items on the test. The sample size for field testing will depend on the intended psychometric analyses; for example, classical item statistics do not require as many student responses as item response theory–estimated parameters. In PAD, just as in conventional test design, data from a well-designed field test are required to do the psychometric analyses described in this section. However, there are various opportunities for gathering data throughout the design process prior to field testing that should be taken.

For example, administering early prototypes of items to relatively small samples of students in settings where students can be observed and interviewed (e.g., cognitive labs) can confirm or help us revise our assumptions about the cognitive processes employed in responding, or the interaction between cognitive demand and item type. We can use this information to update our artifacts that represent the essential elements of PAD: approach to cognition, RPLDs, and/or our task models and difficulty drivers. For computerized assessments, it is strongly advised to frequently observe students navigating through multiple components of the assessment as they are developed, with a keen eye toward how the accessibility features are performing for students with various needs. When time and resources are available, using tools such as eye-tracking equipment can also provide fascinating insights into how students are engaging with the assessment and help improve how the tasks and navigation are designed. At the time of this writing, we are beginning to see testing programs gather input from students on item design as

part of an effort to better represent diversity and inclusion in our assessments. In each of these cases, the data from these opportunities serve as invaluable feedback to the design process and only serve to strengthen the evidentiary argument for the assessment.

Content and Skill Integration

In PAD, the content and skills to be assessed are integrated in the form of claims—for example, a student can classify two-dimensional figures in a hierarchy based on defining attributes, where "classify" is the skill and the content is "two-dimensional figures in a hierarchy based on defining attributes." As discussed in the previous section, these claims are arranged along a progression, in the form of RPLDs, that demonstrate how a student moves within a school year from novice to grade-level proficiency. RPLDs in and of themselves cannot serve as test specifications because test specifications need to consider the number of items, the type of items, and other design features and constraints of the assessment to guide form assembly in a fixed-form assessment or the item selection algorithm in an adaptive testing environment. In a conventional approach, although content and skills may be weighted or balanced in the test specifications (e.g., the 12 items in the geometry domain should have a skills distribution of 30%–40% application and 60%–70% comprehension), the content and skills are not integrated as a claim about what students know and can do. As we argue in the section "Using RPLDs to Guide Assessment Design," this lack of integration does not support the type of scale properties that are optimal for supporting performance standards (also referred to as cut scores) and inferences about what students know and can do that are based on the RPLDs.

Figure 7.10 provides two examples of typical mathematics test specifications for state accountability tests for elementary grades. In each case, the testing program goes on to give descriptions of each component in some detail but falls short of articulating how the content and skills (represented in each case by DOK) should be integrated to support the intended inferences regarding student proficiency. For instance, for the example on the right side of Figure 7.10, the technical manual includes descriptions of both DOK and the mathematical practices, and it indicates that two mathematical practices, *justification and explanation* and *modeling*, result in reportable scores but does not specify guidelines for how the mathematical practices and the DOK levels relate to one another or how they should be integrated into the assessment design. Without a priori hypotheses about how these various design components relate to eliciting the required evidence from students documented in design patterns or other item-writing specifications, these decisions are left to each individual item writer. Regardless of the depth of expertise of any given item writer, it is extremely unlikely that a coherent perspective on the following will occur by happenstance:

- Which item types and item features are best suited to elicit the types of skills implied by DOK 1 versus DOK 2 versus DOK 3?
- What is the best way to elicit evidence of mathematical practices given the item types and manipulatable item features available?
- What are the relationships between DOK and mathematical practices?

Example 1

DOMAIN	MIN.	MAX.
Operations and Algebraic Thinking	29%	38%
Number and Operations—Base 10	18%	22%
Number and Operations—Fractions	27%	31%
Measurement, Data, and Geometry	18%	22%
DOK 1	18%	31%
DOK 2	38%	58%
DOK 3	9%	20%

Example 2

OPERATIONAL ITEMS	NUMBER OF ITEMS
Item Types	
Multiple Choice	21–23
Technology Enhanced	0–6
Constructed Response	3
Reporting Categories	
Progress to Grade Level	19–20
Number and Ops—Base 10	3–5
Number and Ops—Fractions	4–5
Ops & Algebraic Thinking	3–6
Geometry	3–4
Measurement & Data	3–4
Mathematical Practices	
Integration	7–8
Justification/Explanation	3
Modeling	≥7

FIGURE 7.10

Conventional Test Specifications With Separate Considerations of Components

Note. DOK = Depth of Knowledge.

- What is the optimal way to pair content and skills (or, in these examples, DOK level) given the intended inferences about student proficiency?

In their chapter on the role of cognitive models in large-scale educational test development, Huff et al. (2017) contended that with PAD,

> all assumptions, especially those regarding student cognition, are brought to the foreground and examined. As a result, when the cognitive model is selected a priori and is made explicit, there is the opportunity to ensure alignment of the cognitive model with the intended purpose and use of the assessment as well as with all other design, scoring, reporting and interpretation decisions to yield evidentiary coherence. (p. 401)

In contrast, assessment programs that employ PAD in some way tend to have test specifications that integrate the content and skills to be measured in ways that make transparent the assessment designers' perspective on what is valued in the domain. The NAEP Technology and Engineering Literacy assessment specifications are a great example of this approach. Figure 7.11 presents a snapshot of a rich set of assessment specifications that integrate the content and skills that are valued in the classroom and discipline (National Assessment Governing Board, 2018). Note that each content area (columns) is crossed with the practice (rows), which gives rise to a differentiation in the

Classification of Types of Assessment Targets in the Three Major Assessment Areas According to the Practices for Technology and Engineering Literacy

	Technology and Society	Design and Systems	Information and Communication Technology
Understanding Technological Principles	**Analyze** advantages and disadvantages of an existing Technology **Explain** costs and benefits **Compare** effects of two Technologies on individuals **Propose** solutions and alternatives **Predict** consequences of a Technology **Select** among alternatives	**Describe** features of a system or Process **Identify** examples of a system or Process **Explain** the properties of different Materials that determine which is suitable to use for a given Application or product **Analyze** a need **Classify** the elements of a system	**Describe** features and functions of ICT tools **Explain** how parts of a whole Interact **Analyze** and compare relevant Features **Critique** a process or outcome **Evaluate** examples of effective Resolution of opposing points of View **Justify** tool choice for a given Purpose
Developing Solutions and Achieving Goals	**Select** appropriate technology to Solve a societal problem **Develop** a plan to investigate an Issue **Gather and Organize** data and Information **Analyze and Compare** advantages and disadvantages of a proposed Solution **Investigate** environmental and Economic impacts of a proposed Solution **Evaluate** trade-offs and impacts of a proposed solution	**Design and Build** a product Using appropriate processess and Materials **Develop** forecasting techniques **Construct and Test** a model or Prototype **Produce** an alternative design or Product **Evaluate** trade-offs **Determine** how to meet a need by choosing resources required to Meet or satisfy that need **Plan** for durability **Troubleshoot** malfunctions	**Select and Use** appropriate tools to Achieve a goal **Search** media and digital resources **Evaluate** credibility and solutions **Propose and Implement** strategies **Predict** outcomes of a proposed Approach **Plan** research and presentations **Organize** data and information **Transform** from one Representational form to another **Conduct** experiments using digital Tools and simulations

FIGURE 7.11
Snapshot of Integrated National Assessment of Educational Progress Technology and Engineering Literacy Specifications
Note. ICT = Information and Communication Technology.

targets of measurement that exist at the intersection of content and skills, presenting the item writer with a more detailed set of expectations to guide their work.

Using RPLDs to Guide Assessment Design

The second area of integration that is a hallmark of PAD is using RPLDs as an essential element of assessment design. Recent reviews of testing programs' use of PAD (Huff et al., 2017) indicate that the use of RPLDs to guide design is still nascent at best, although using them has been long called for in designing assessments for which the primary purpose is to reliably classify students into a proficiency level and support valid inferences about what they know and can do (Bejar et al., 2007; Huff & Plake, 2010a; Luecht, 2013, 2019; Perie & Huff, 2016).

In conventional approaches to assessment design, articulation of PLDs of any sort does not occur during the design process; rather, the development of PLDs is considered the first step of the standard-setting methodology, which occurs after the

first test administration. In contrast, in PAD the items are designed to elicit evidence for a particular performance level (or, more specifically, for a particular claim about students where the claims are arranged along a performance continuum as expressed in the RPLDs).

This intentional, a priori match of an item to a specific performance level has several benefits. First, the claims about what students know and can do in each performance level are inferences that require an evidential argument that can be evaluated. To have reasonable confidence that inferences about what students know and can do are valid, this evidential argument must be compelling. Designing items from the outset that target specific performance levels seems to be an inherently plausible, commonsensical approach. The alternative argument—that items designed *absent* RPLD inputs will yield valid evidence helpful for classifying students based on what they know and can do—seems whimsical at best. A second, related benefit is that when items are designed to target specific RPLDs, we are essentially engineering the desired scale properties from the beginning, creating a coherence and integration between item design and psychometrics that is fragmented in conventional approaches to test design. Third, when items are targeted to specific claims about students organized as RPLDs, there is an opportunity to engineer cut scores rather than have standard setting be a completely ad hoc process (Huff & Plake, 2010b; Lewis & Cook, 2020).

Resistance to targeting items to specific performance levels is typically rooted in the observation that the intended progression of items does not turn out perfectly when reviewing the observed difficulty statistics for the item. For example, suppose there is an assessment that is intended to have three performance levels that are represented by the item response theory–based scale scores, as shown in the first column of Figure 7.12. Items were designed to align to performance levels, as shown in the second column. However, after data were collected from students in initial field trials and the items were analyzed and placed on the scale according to their estimated difficulty parameters (third column), there was a mismatch for many of the items. Some would use these mismatches as evidence that attempting to target items to specific performance levels is too imperfect of a science to be of utility. However, the mismatch of observed difficulty to intended difficulty is not a new problem—conventionally developed norm-referenced tests overdevelop item pools for precisely this reason, where crafting intended scale properties is mostly achieved through careful selection from an item pool after field testing, rather than designing the fit of items to scale intentionally from the beginning. But others argue—including the authors—that the process of targeting items to specific performance levels, reviewing data from field tests, hypothesizing on the cause of the mismatches, and revising items when needed—is simply best practice in the spirit of continuous improvement and iterative design (Kaliski et al., 2011; Lewis & Cook, 2020; Luecht, 2013, 2019; Schneider et al., 2013).

In the spirit of PAD, assessment designers have advocated seeing these mismatches as an opportunity for discussion and learning rather than dismissal of the entire process. Indeed, they contend that these discussions should occur in relation to all items,

not just the ones that are mismatched (Huff & Ferrara, 2010; Lewis & Cook, 2020). The following three questions should guide the evaluation of items: (a) Does the item placement along the scale support or detract from the assumptions as articulated in the domain model essential elements (approach to cognition and RPLDs) and the design patterns (task features, difficulty drivers)? (b) Do any assumptions or design features need to be revised? (c) What can we assume about opportunity to learn (OTL), and how does that influence the data for the items? For items that are intended to be rather simple in complexity and the statistics on the item support this intention, like Item 1 from Figure 7.12, it is usually assumed that most students have had OTL to learn the knowledge and skills tested by the item and that students in the *Basic* category have some barrier to OTL (e.g., do not have the prerequisite skills to access this grade-level content). These same types of discussions should happen with items at the other end of the spectrum, like Item 9 from Figure 7.12. The item was intended to be complex and the item location along the scale seems to support the assumptions that went into the RPLD and item design. However, the design team reviewing the item and resulting statistics should not discount the counterhypothesis that the item is complex as the result of some flaw (e.g., construct-irrelevant variance) or that our assumptions about the complexity of the target of measurement are flawed and the main reason that the item statistics are showing the item as hard is that very few students have had OTL. The role of OTL is especially critical to consider for items that are mismatched: All of the assumptions that go into the item design can be supported by research and expert judgment, the item can be free of flaws, and the item can still locate on the scale in an unexpected way because of a lack of OTL (making the item seem more complex than it is had students had the OTL for the targeted knowledge and skill), widespread

RPLD and Scale Score Range	Intended RPLD	Observed Placement on Scale
Basic 350–450	Item 1	Item 1
	Item 2	Item 4
	Item 3	Item 3
Proficienct 451–550	Item 4	Item 2
	Item 5	Item 5
	Item 6	Item 8
Advanced 551–650	Item 7	Item 6
	Item 8	Item 7
	Item 9	Item 9

FIGURE 7.12
Hypothetical Comparison of Intended Performance Levels and Item Scale Placement
Note. RPLD = range performance-level descriptor.

exceptional instruction in a particular area (the targeted content and skill is very complex and sophisticated, but most students are getting the item correct because they have had effective instruction and OTL), and/or more nefarious reasons such as item exposure or cheating (making complex items seem easy).

The mismatch between items and intended performance level needs more attention because understanding these mismatches helps us make better assessments and, in turn, may help us better understand—and improve—the assumptions about the construct of interest, student cognition, the underlying performance continuum as expressed by the RPLDs. It is in this process that our assumptions are either confirmed or disconfirmed and thus revised. This is the messy business of reasoning from imperfect evidence that is at the heart of PAD.

Using PAD to Engineer Desired Scale Properties

Tests for which the primary purpose is to provide norm-referenced information, such as a percentile rank, or tests for which the primary validity evidence is the strength of the prediction from a score to a criterion, such as grade point average, do not require sophisticated psychometric specifications. Typically, for these kinds of tests, psychometricians give very little direction to item writers in advance other than general advice to make sure there is a range of difficulty in items. It is assumed that high-quality items will produce an appropriate range of item discrimination and very few items will be flagged for differential item functioning. The goal of the psychometrician for these kinds of tests is to keep reliability high and maintain the unidimensional scale over time as new items are added and the population shifts. Norms and predictions are monitored and updated on a routine but not annual basis. The scale is reset as infrequently as possible because maintaining historical trend is of utmost importance to the score user. In this context, the psychometric team will review test specifications upon development or revision in conjunction with the item-writing team to ensure that there are enough items of various characteristics (e.g., algebra or geometry, multiple-choice or short constructed response) to accomplish these goals. However, in conventional testing programs, psychometricians do not become heavily involved in the design of the assessment or items until there are field test data to analyze. As Luecht (2019, p. 8) lamented,

> The fundamental dilemma seems to stem from the way that we design and implement operational testing programs and ultimately develop our score scales. Most test forms and the score scales are usually constructed in siloed procedures, carried out by different groups within an organization doing different things at different times without extensive coordination nor careful integration when it comes to validity and score use. [Item writers] tend not to worry about the statistical properties of the scale, psychometricians often view test content as an annoyance that gets in the way of optimizing certain statistical scale properties, and end users must then rely on PLDs and standard setting and ad hoc score reporting techniques to build some semblance of meaning into the score scales.

For tests that need to support inferences about student proficiency in regard to a set of learning standards, we argue that psychometricians and item writers must be close collaborators from the beginning and throughout the design process, as part of an interdisciplinary team with representatives from other, related fields, such as but not limited to classroom teachers and learning scientists, long before test specifications are developed or items are field tested. Through this collaboration, the intended scale properties can play a central role in the design process. Specifically, assessments that will require performance standards—or cut scores—so that each student can be classified into a proficiency level require sufficient measurement information to distinguish students who are, for example, *Basic* versus *Proficient* versus *Advanced*. For this reason, item writers need to change their perspective from writing items that span a wide range of difficulty to targeting the items to distinguish between performance levels, which requires the RPLDs to inform assessment design and item writing. The criticality of doing so must be understood by psychometricians and item writers alike for the scale to support a successful standard setting (Huff & Plake, 2010b; Lewis & Cook, 2020; Luecht, 2013, 2019) and for the resulting inferences to have compelling evidentiary argument that leads to validation of the inferences about what students in each performance level know and can do. The intentional relationships among our research-based assumptions about student cognition, the latent proficiency continuum as expressed in the RPLDs, task features, difficulty drivers, and the resultant scale properties of the assessment should result in a strong body of evidence to support valid inferences about what students know and can do.

One hypothetical example of what an integrated set of content and skills to be assessed and psychometric specifications could look like is provided by Luecht (2019) in Figure 7.13. In this example, Luecht used the NGSS to illustrate that when appropriately engineered, design patterns (see the term *task model* in Figure 7.13) will produce measurement information that is located on a deliberate range of the scale to support intended inferences. This can only happen when RPLDs—composed of claims about student proficiency that integrate skills and content and reflect a progression of student cognition—and the desired scale properties are understood by the full interdisciplinary team to be complementary components and are used as the basis of the assessment design from beginning to end.

SUMMARY

PAD helps us integrate and create coherence where conventional approaches to assessment design are fragmented. Three areas of integration were discussed in this section: first, test specifications that integrate content and skills in meaningful ways to reflect the hypothesized learning trajectory articulated in the RPLDs and serve as the basis for the eventual inferences about what students know and can do; second, the use of RPLDs as the basis for iterative item design and development; and third, the integration of item

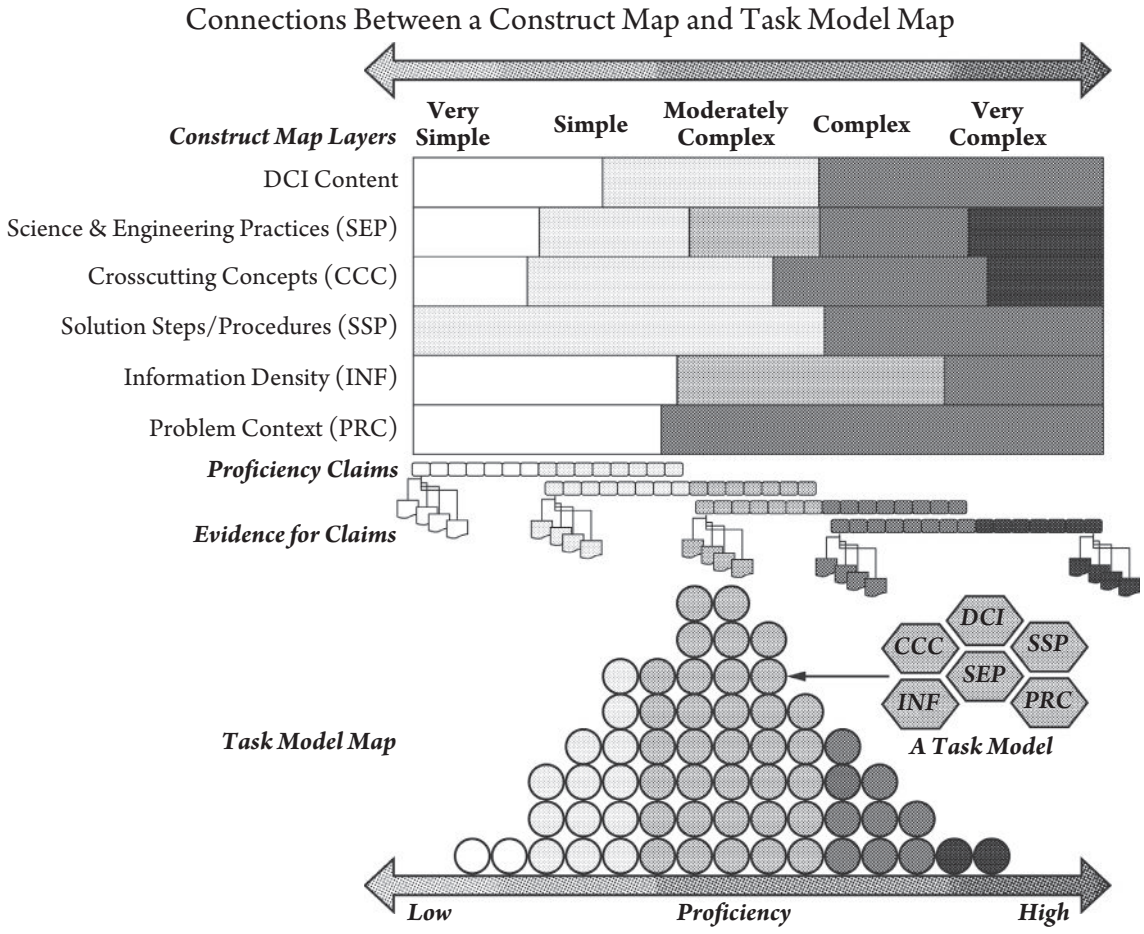

FIGURE 7.13

Generating Performance-Level Descriptors for the Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores Partnership

Note. From "Strengthening Claims-Based Interpretations and Uses of Local and Large-Scale Science Assessment Scores (SCILLSS): The Role of Performance Level Descriptors for Establishing Meaningful and Useful Reporting Scales in a Principled Design Approach (White Paper)" by R. M. Luecht, January 2019, Nebraska Department of Education. https://www.scillsspartners.org/wp-content/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf

design and scale design. These three related levels of integration require collaboration across an interdisciplinary team from the beginning and throughout the iterative design and development process, especially by those responsible for designing and developing the items and those responsible for designing and maintaining the scale through psychometrics. It is this same partnership that is required to articulate a robust inferential argument that is supported by evidence and supports valid interpretations of what students know and can do.

The integrations discussed in this section have profound impact on how we think about the inferential and validation arguments. With PAD, we have the opportunity to develop a validation argument that is shaped by the deeper notion of coherence, rather than the typical superficial alignment, and includes evidence to support our intended

interpretations of student performance that is generated from the beginning of the design process, not post hoc validation studies.

THE EVIDENTIARY ARGUMENT IN PAD

One hallmark of PAD is that the evidentiary argument (which is defined as the evidence to support the inferential argument) and the validation argument (which is defined as an evaluation of how compelling the evidence supporting the inferential argument is) are central components of the iterative design endeavor, rather than a series of post hoc analytical exercises to be documented in a technical manual. For the purposes of this chapter, we posit that:

1. A Kanesian perspective on assessment validation (Kane, 2006, 2013) complements both a theory of action framework and the PAD framework.
2. A *theory of action* (TOA) is required to frame the evidentiary and validation arguments for assessment in the larger educational context.
3. The evidentiary argument and the validation argument are composed of a series of claims and evidence that are hierarchical, nested, and represented in many forms and in many grain sizes. For example, the *evidence model* in ECD is an evidentiary argument for specific targets of measurement, but only a subcomponent of the larger evidentiary argument supporting score interpretation, which in turn is only a subcomponent of the larger evidentiary argument supporting the effectiveness of score use as framed by the TOA.
4. The educational assessment industry needs to hold itself accountable for the decisions made throughout the assessment design process that either support or hinder the intended inferences about what students know and can do through the production of *procedural validity evidence for assessment design*.
5. Given this context, in PAD, many post hoc "validity studies" become obsolete. We use post hoc alignment studies as the prime example.

We discuss each of these propositions in turn in the following section.

Assessment Validation

As discussed previously, Toulmin's (1958, 2006) thinking on the use of practical arguments, rather than formal logic, played an influential role in the evolution of PAD. The extension of the use of practical arguments to assessment validation processes has been an integral part of PAD from the outset. Kane (2013) described an argument-based approach to validation as requiring that

> the claims based on the test scores be outlined as an argument that specifies the inferences and supporting assumptions needed to get from test responses to score-based interpretations and uses. Validation then can be thought of as an evaluation of the coherence and completeness of this interpretation/use argument and of the plausibility of its inferences and assumptions. (p. 1)

In PAD, we think of the evidentiary argument as the evidence used to support the inferential chain of reasoning threaded throughout the design components of the assessment. What evidence do I have to support that the approach to cognition is appropriate for this domain? These grade levels? What evidence do I have that the RPLDs follow a reasonable progression of student learning from novice to mastery within a given grade and domain? What evidence do I have to support that the design patterns and resulting items will yield evidence of where along the latent proficiency continuum a student most likely resides? What evidence do I have that the resulting performance-level classifications support intended inferences given the stated purpose and use of the assessment? The answers to these questions—and many more like them—constitute the evidentiary argument. The validation argument is an evaluation of how compelling the evidentiary argument is. It is incumbent on the test publishers to provide the evidentiary argument. The potential test users determine the persuasiveness of the evidentiary argument by their decision to use or not use the test for its stated purpose and use. However, that decision does not constitute a "validation" of the test. Validation would require an external review of the evidentiary argument that evaluates the documentation of the design process and the articulated answers to all of the questions that arise through that process, as well as a determination of whether the evidentiary argument is, indeed, compelling and coherent.

Theory of Action

Since 1992, states have been mandated by the U.S. Department of Education to test, on an annual basis, certain subjects in certain grades for various accountability purposes and with some promise that these assessment results can also inform decisions at various levels of the educational system (federal, state, district, building, classroom). For example, when the No Child Left Behind Act became federal law in 2002 (No Child Left Behind Act, 2002) and extended the requirement from the previous law (Improving America's Schools Act, 1994) that state assessment results be disaggregated by various subgroups, states used those data to make decisions about how to allocate funding for curricular supports, professional development, and other services. Districts also administer a variety of assessments throughout the school year for various purposes and uses, as do principals at the school level and classroom teachers. When thought through with purpose, these assessment endeavors complement each other in a balanced assessment system (Marion et al., 2019; Perie et al., 2009). When not coordinated as a system, this results in a proliferation of redundant testing that can have many negative, unintended consequences on teaching and learning. (See Ho & Polikoff, this volume, for additional discussion of test-based accountability in K–12 education.)

Bejar (2010), building on work from Bejar et al. (2007) and Mislevy and Haertel (2006), suggested that when PAD is employed, it is incumbent on the designers to consider both construct representation and convergent and discriminant validity evidence. Bejar (2010) highlighted the latter

> because test scores, certainly those of high visibility, do not live in isolation and have consequences, intended and unintended, for students and their parents. Such consequences can be thought of as part of the evidence for and against a particular assessment. (p. 381)

Bennett (2015) further extended this notion of examining consequences by postulating that in a teaching and learning context, assessments are a form of treatment and the intended effects of the assessment must be examined through the lens of efficacy or TOA.

Figure 7.14 provides a sample TOA for an interim assessment for which the primary purpose and use are to help teachers identify gaps in student learning in a particular domain of interest (e.g., fourth-grade learning standards for measurement and geometry) and to support changes in instructional practice with the intended outcome of improving student learning. From this perspective, both the evidentiary argument and the validation argument are subsumed by a broader TOA for the role that the interim assessment plays in teaching and learning.

In this example, evidence to support fidelity of implementation is a key component of the TOA. For example, although the score reports may be optimally designed to inform instructional actions, without the appropriate support systems in place the reports may not be used as intended. One can imagine several systemic actions that would need to be evaluated: Was adequate professional development provided to

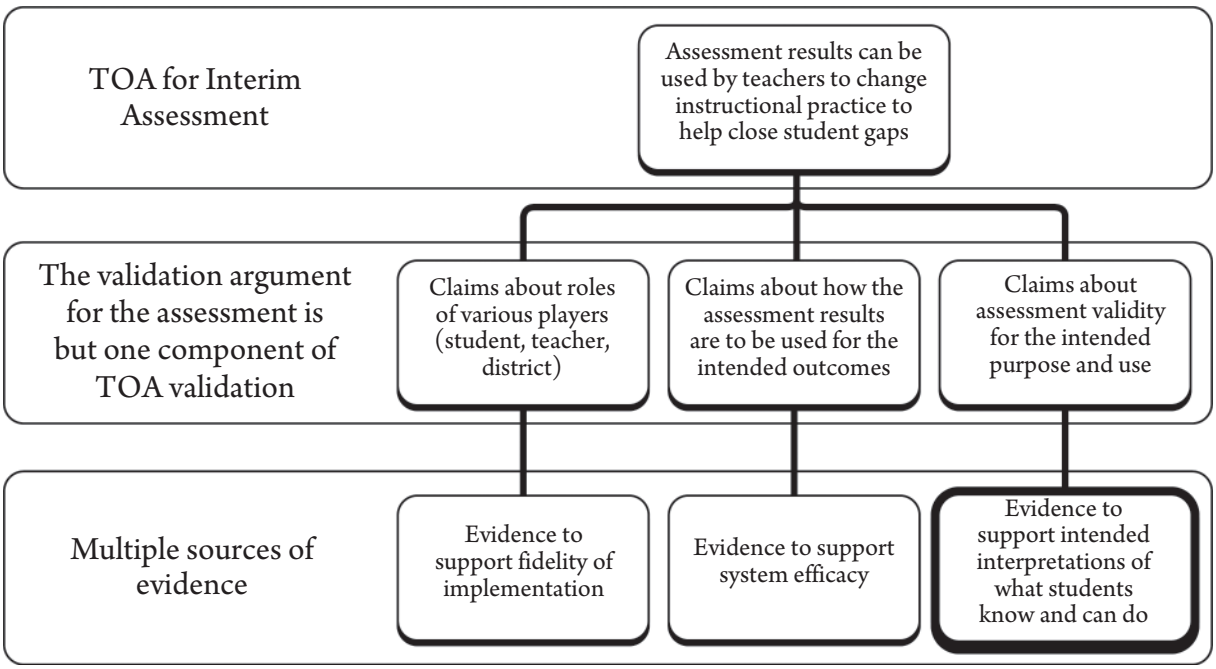

FIGURE 7.14

Sample Theory of Action for Interim Assessment

Note. TOA = Theory of Action.

support appropriate interpretation and use of the reports? Were teachers given timely access to the reports? Were teachers provided time to review the reports and plan accordingly? Do teachers have access to curricular and instructional supports that cohere with the assessment results? When not implemented with fidelity, it is unlikely that an assessment designed to adjust instructional practice will result in efficacious student learning gains.

Research that examines the impact of the interim assessment as a "treatment" should occur to determine whether the actions made based on the score report result in improved achievement for students (Brookhart, 2009; Nichols et al., 2009; Shepard, 2009). There is a growing body of research suggesting that teachers often struggle to determine next instructional actions to take (Heritage et al., 2009; Ruiz-Primo et al., 2010; Schneider & Andrade, 2013; Schneider & Gowan, 2013; Schneider & Meyer, 2012). In light of this finding, if a testing program's claim that the intended purpose and use of the assessment is to inform instruction, this must be supported by assessment designers' documented work with an interdisciplinary team—expert teachers, learning scientists, score report designers, to name a few—to develop instructional recommendations. To return to Figure 7.14, score interpretations must be supported by evidence that supports their accuracy (evidentiary argument) and instructional recommendations should be supported by evidence that demonstrates that faithful implementation of the recommendations results in student learning gains (efficacy research).

Most testing programs focus on the evidence to support score interpretation in the lower right box of Figure 7.14 to the detriment of other sources of evidence. Gathering and documenting evidence to support score interpretation is usually treated rather mechanically, as demonstrated by the routine structure of most technical manuals that include a predictable series of psychometric analyses. When PAD is used, we argue that a rich and coherent evidentiary argument is articulated by the very nature of the design process itself.

Claims and Evidence

Assessments are a type of evidentiary argument (Kane, 1992; Messick, 1989; Mislevy, 1994). As our claims about what students know and can do become more complex and the uses of assessment results become more varied and complex, the burden on our evidentiary argument increases. Simultaneously, as educational tests come under greater scrutiny, it is incumbent on test makers to ensure that the evidentiary argument supporting the inferences we are making about students is both clear and compelling. There are at least two ways that PAD differs from conventional test development that help tremendously with creating a clear, compelling evidentiary argument. First, the inferential argument is defined up-front, used as the basis of design, documented clearly throughout the design process, and updated during assessment development; this documentation serves as both design tools (i.e., the essential elements of PAD) and the evidentiary argument that supports the inferential argument. Second, one way that evidence is conceptualized in PAD is the

observable manifestation of what the student knows and can do, and the articulation of that which is observable is used as a design element throughout the assessment development process. Collecting and curating these elements of the evidentiary argument throughout the design process contrasts with conventional approaches to test development, in which "evidence" is narrowly defined as the collection of empirical data that can only happen when design is complete and scores are analyzed.

Under PAD, the term *evidence* is used in different contexts to reference various aspects of the larger body of evidence. Evidence is used broadly to refer to the evidentiary argument that captures the comprehensive set of claims that begins with the definition of the construct and ends with the use of the test score information to accomplish the intended purpose of the assessment. During design, development, and administration, these claims are supported by the collection and documentation of design process artifacts, including but not limited to the essential elements and empirical data. Psychometric evidence is collected to support claims that items are appropriately designed and scored to locate students along the performance continuum as articulated by the RPLDs. In addition, evidence also refers to the data, warrant, and backing supporting a claim of the TOA. Table 7.3 gives examples of how claims and evidence are used in nested levels within PAD as design features as well as components of the evidentiary argument.

In Figure 7.15, we show the hierarchical and nested nature of the claims that constitute the evidentiary argument for an examination program designed to award college credit and placement.

Procedural Validity Evidence

PAD, as described in the previous section, evolved from a rich history of the role of evidentiary reasoning in assessment; a natural consequence of engaging in PAD is the documentation and use of several layers of the evidentiary argument in the form of the various artifacts that represent the essential elements of PAD, for example, prioritized knowledge and skills in the domain (domain analysis), the approach to cognition and RPLDs (domain model), and the task features and difficulty drivers (design patterns). Nichols et al. (2017) posited that to achieve the intended purpose and use of an interim or summative assessment in the educational system, coherence must exist in the design process in at least two ways: (a) All design elements must be informed by theories of learning and cognition, and (b) procedural validity evidence must be collected to demonstrate that the assessment designers followed the process and used the design elements as intended. If the intended process was followed, then the procedural validity evidence will be natural artifacts of that process.

PAD starts with an assessment design plan that explicates in detail the processes that will be followed during design, the points of iteration, the artifacts that will be produced, and the key attributes of those artifacts. In this way, procedural validity evidence for the claim "The intended assessment design process was followed" can be collected and evaluated. Such a notion may seem novel in assessment design, but it is not. We

Table 7.3 Claims and Evidence in Principled Assessment Design

Grain Size	Claim	Example Evidence
Smallest: A single item with a particular *evidentiary focus* (i.e., target of measurement)	Student can distinguish among levels of abstraction in conflicting historical information	Task stimuli (historical text, map, and/or graphic) includes various levels of abstraction. Task prompt focuses student on distinguishing among levels of abstraction
Small: A single learning standard	Student can evaluate conflicting historical information for particular historical period and focus (e.g., pre-Columbian migration patterns)	Observable evidence has four components: • Recognition and response to conflicting information • Clear, comprehensive thesis • Significant depth and quantity of evidence to support thesis, including main concept and supporting details • Recognition of different levels of abstraction in historical information
Medium: A primary design component for the assessment (e.g., RPLDs or design patterns)	The RPLDs represent a research-based progression of proficiency that are useful both for assessment design and for making inferences about what students know and can do	Literature review on how students learn and build knowledge and skills in the domain of interest; data (survey, focus groups) that indicate educators are making appropriate inferences about what students know and can do from the PLDs; procedural validity evidence from the design pattern development and item-writing process that PLDs are being used as intended in design
Medium: One of the five sources of validity evidence cited in the *Standards* (AERA et al., 2014)	Students employ intended cognitive processes when responding to assessment items	Research report from a think-aloud study that indicates students are using the intended cognitive processes at each performance level
Large: Evidence that one or more claims in the TOA are supported	When teachers use the assessment results to drive instruction, students achieve higher gains in learning than when teachers do not	Quasi-experimental design study showing respectable effect sizes for claim that when teachers use the results of the assessment to drive instruction, students have higher gains than when assessment results are not used

Note. PLD = performance-level descriptor; RPLD = range performance-level descriptor; TOA = theory of action.

routinely articulate our intended methodologies and our rationales for every psychometric aspect of the test-making endeavor: establishing and maintaining the scale over time, equating or linking forms (or monitoring item drift), item analyses, standard setting, and postoperational validation studies (e.g., dimensionality analysis, predictive studies). These plans are meticulously detailed by the assessment provider, reviewed by independent technical advisors who provide feedback, and then reviewed and approved by the assessment client (e.g., the district or state). The assessment providers are held responsible for incorporating feedback from the advisors and client and then are held accountable for executing the psychometric plans and presenting detailed evidence of said execution. In addition, there are widely held expectations throughout the industry

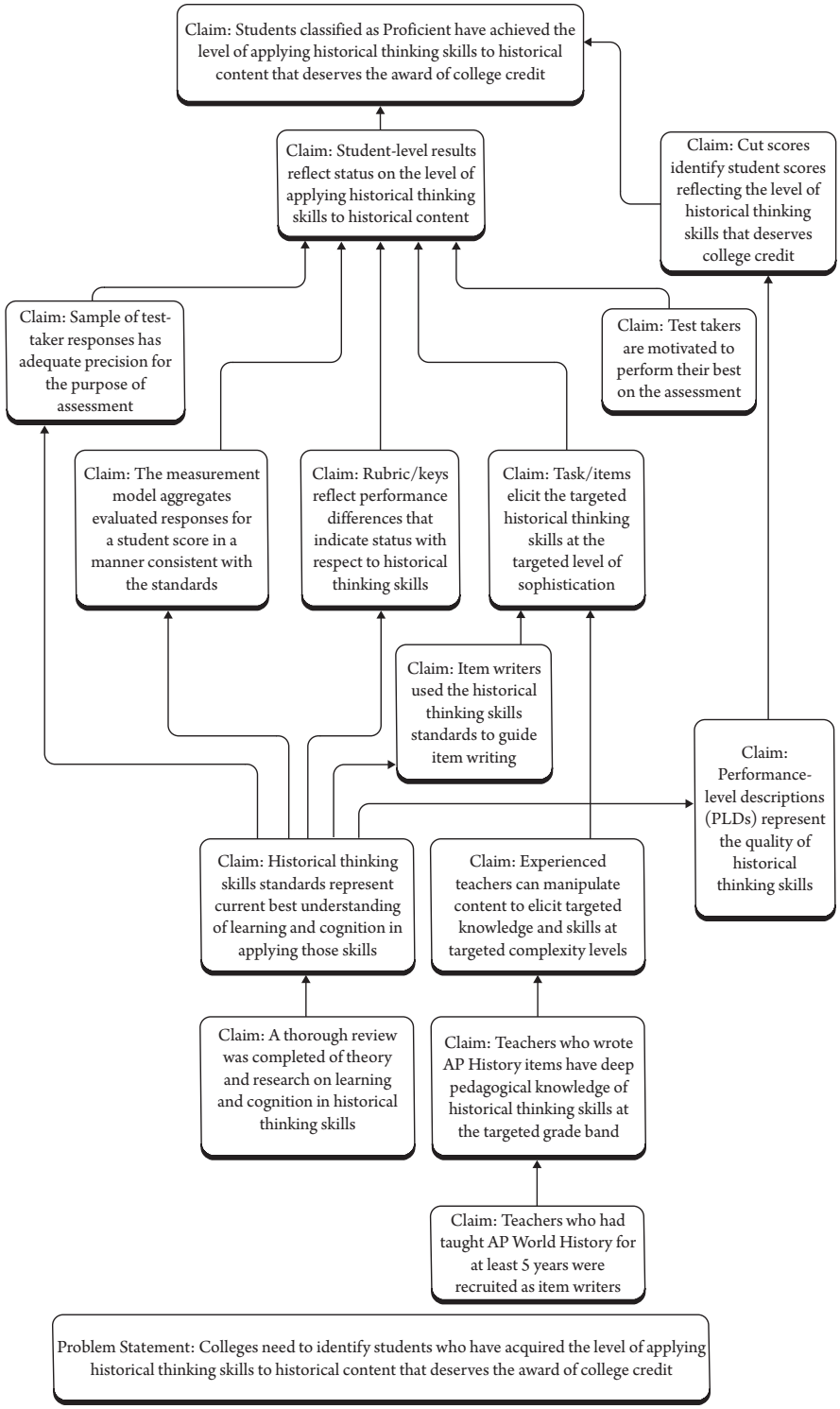

FIGURE 7.15
The Nested and Hierarchical Nature of Claims

on what constitutes acceptable and thorough documentation of these psychometric methods and analyses, which for federally mandated state summative tests are, in fact, externally audited via peer review. The whole of assessment design and development should be held to the same rigorous expectations of making a plan, executing it, and providing evidence that it was executed as intended.

An assessment design plan outlines the design and development process in detail, as well as what evidence will be produced that demonstrates that the plan was executed. Given that most of design and development relies on human judgment, assessment designers should be held accountable for providing procedural validity evidence. When PAD is engaged, procedural validity evidence is a natural by-product of the process as documentation of the assumptions and rationales undergirding each decision is used as design tools (e.g., the approach to cognition, design patterns). It is also helpful to be explicit in the plan on the role and nature of iteration, especially regarding the use of field test data to refine RPLDs and design patterns.

Table 7.4 contains a proposed set of validation criteria for PLD development expanded to include criteria for design patterns and item development. Huff and Plake (2010b) adapted these criteria from what are typically required of the standard-setting process. There is no reason that we should not apply the same high expectations to assessment design and development.

Hendrickson et al. (2013) offered two sets of checklists for use in the PAD endeavor. One is a set of criteria that can be used to evaluate design pattern quality. The criteria are based on assumptions that the items are written to discern between increasing levels of performance as articulated in the RPLDs and that there will be iteration of either or both RPLDs and design patterns after field testing. The second example from Hendrickson et al. (2013) is a checklist regarding iteration for PAD design components that relies heavily on iteration and consensus across various stakeholders on an interdisciplinary team. How do we know when we are done and can move forward to the next phase?

A final example in Figure 7.16 represents an item specification checklist to help ensure that the item writer is keeping critical design elements in mind. Notice how the checklist requires the item writer to indicate whether the item is to discern between students who are at the lower end of the proficiency continuum (Performance Levels 1 and 2: identify) or between students who are at the upper end of the proficiency continuum (Performance Levels 3 and 4: analysis) and how those items have different characteristics related to student cognition. It is easy to see from this example how a design tool can also serve as procedural validity evidence and evidence to support an inferential argument that makes explicit the relationships among approach to cognition, RPLDs, design patterns, items, and inferences about what students know and can do.

Alignment

Alignment of test specifications and item content to the learning standards is a key tenet of content validity for educational assessments. In conventional approaches to test design, alignment is conducted after item writing is complete. We contend that when

Table 7.4 Validation Framework and Criteria for Performance-Level Descriptors, Design Patterns, and Item Development

Category of Evidence	Criterion	Potential Evidence for PLDs
Procedural	Care in selecting panelists	Qualification and representation of panelists to support claims panel has expertise in how students learn and progress in domain and what it looks like as students reach proficiency and beyond
	Justification for PLD development framework	The approach chosen should be justified and the learning science paradigm made explicit
	Panelist training	Surveys indicate panelists understand the framework and have been sufficiently trained
	Appropriate contribution from panelists	Panelists' contribution, discussions, points of consensus, and compromise should be documented; panelist surveys
	Proper implementation	Documentation of implementation compared to workshop design: agenda, panelist surveys, PLD creation development templates; training framework
	Panelist confidence	Survey at end of panelists' confidence in process, quality, and successful implementation of development framework
Internal	Sufficient interpanelist consistency	Points of disagreement among panelists should be addressed and effectively moderated to achieve consensus or compromise
	Decreasing variability across rounds	Panelists' judgments should converge throughout the PLD development process via consensus building
	Consistency across independent panels	PLDs from two independent panels should have multiple points of consistency and few points of difference; document through qualitative comparison
	Consistency across panel subgroups	Reasonable panelist characteristics (e.g., teacher vs. researcher, assignment to different discussion groups during PLD session) should not impact
External	Expert reviews	Expert reviews should confirm evidence from learning sciences is found within PLDs; expert reviews of PLDs should confirm utility for item writing
	Teacher review	Resulting PLDs should be reviewed for interpretability by a separate panel of teachers or through a public review
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?
Procedural	Care in selecting designers	Qualification and representation of panelists to support claims that panel has expertise in how students learn and item design
	Justification for task model development framework	The approach chosen should be justified and the grain size of and intended design patters for item types made explicit in cases where design patterns are not created for all item types
	Designer training	Surveys indicate that designers understand the intended goals and outcomes of design patterns and have been sufficiently trained

Category of Evidence	Criterion	Potential Evidence for PLDs
	Appropriate contribution from designers	Designer contribution, discussions, points of consensus, and compromise should be documented; designer surveys
	Proper implementation	Documentation of implementation compared to planned design: agenda, surveys, creation of design patterns; training framework
	Designer confidence	Survey to assess designers' confidence in process, quality, and successful implementation of development framework
Internal	Sufficient consistency	Points of disagreement among designers should be addressed and effectively moderated to achieve consensus or compromise
	Increased confidence across iterations	Iterative feedback should increase in perceived utility for consensus building
	Consistency across subgroups	Designer characteristics (e.g., teacher vs. researcher, assignment to different grade levels) should not impact quality of outcome
External	Expert reviews	Expert reviews should confirm that evidence from learning sciences is found within design patterns; expert reviews of design patterns should confirm utility for item writing
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?
Procedural	Care in selecting item writers	Qualification and representation of item writers to support claims item writers have expertise in the item writing and subject area and grade level assigned
	Item-writer training	Item writers are trained on the use of design patterns and show evidence of being sufficiently trained (e.g., can diagnose features not included in item that should be for a specific achievement-level target) before being allowed to develop items independently
	Proper implementation	Artifacts of training implementation including training framework, practice materials, feedback, and training recordings are collected
	Item-writer confidence	Survey at end of item-writer training to show confidence in process, quality, and readiness to move forward
Internal	Sufficient intraitem writer consistency	Item writer shows evidence of developing and/or identifying intended item features to match PLDs, including appropriate item types for descriptors, and receives feedback when needed
	Increased confidence across iterations	Item writers' feedback on perceived utility should increase as process is implemented
External	Expert reviews	Expert reviews should confirm evidence of training is sufficiently robust; expert reviews of prototype tasks should confirm utility for item-writer training
	Reasonableness	Overall, is the process reasonable, defensible, and free of fatal flaws?

Note. PLD = performance-level descriptor. Adapted from "Innovations in Setting Performance Standards for K–12 Test-Based Accountability," by K. Huff and B. Plake, 2010, *Measurement: Interdisciplinary Research and Perspectives, 8*(2), 130–144.

Grade 6- RL.3

Describe how a particular story's or drama's plot unfolds in a series of episodes as well as how the characters respond or change as the plot moves toward a resolution

Analyze how and why individuals, events, or ideas develop and interact over the course of a text.

Central aspect:
- *Items measure students' ability to analyze how characters change as the plot moves toward resolution. Elements may also include how a plot unfolds in episodes*

Item Measuring CCLS RL.3	Yes/No	If "No," Explain or Describe
Measures central aspect: (PL 1–2) The item requires identification of the change in a character as plot unfolds. Possible stems may include: **Stem:** How does character X change in lines XX–XX? **Stem:** Which of the following best describes the change in character X in lines XX–XX? **Stem:** What does line X reveal about a character? OR Measures supporting aspect: (PL 3–4) The item requires analysis of change or shift in plot Possible stems may include: **Stem:** The change/shift in lines XX-XX develops the plot by **Stem:** Which lines from the story show the character's change from X to Y?		
The item stem does not reveal: • the interaction of elements • the key change/development of characters Unless the interaction is identified in the stem intentionally		
The analysis in the item is supported by the text (i.e., there is development of story elements)		
The item requires students to comprehend the majority of the passage to answer the item correctly		
THE ITEM MEASURES THIS STANDARD		

FIGURE 7.16

Item Development Checklist Example

Note. From "Large-Scale Standards-Based Assessments of Educational Achievement," by K. Huff, Z. Warner, and J. Schweid, in A. A. Rupp & J. P. Leighton, (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 397–426), 2017, Wiley–Blackwell.

PAD is employed, post hoc alignment studies are largely obsolete because the evidence of alignment is produced as part of the design process.

For example, in PAD, items are designed to measure a specific performance level of a learning standard in the RPLDs, and this level is documented. It is best practice that after items are developed, they are reviewed by parties external to the design team, such as some combination of classroom teachers, content domain experts, pedagogy experts, accessibility experts, and experts in cultural and linguistic responsiveness. As part of their review process, the experts could indicate with which RPLD level the item best aligns. These data could be analyzed for interrater reliability. This alignment evidence can also serve as a way to identify areas where more work may be needed; for example, low interrater reliability estimates for particular items can spark the same types of discussions among the design team as when psychometric analyses reveal where along the scale the item locates and whether that location is as intended.

Post hoc alignment studies may still have a defensible role in instances where assessment items were originally designed to meet one set of learning standards, and because of cost and/or time constraints, the items need to be examined for alignment to a revised or different set of learning standards. This is especially true in the state testing context because learning standards tend to change frequently, in either superficial or meaningful ways, as a consequence of leadership changes and political reasons rather than in response to new learning science research. Most states require alignment studies that show that items designed for a previous or different set of learning standards can be used with confidence to assess their new or revised learning standards because assessment design is expensive, and in the K–12 context, the cost ultimately falls on the shoulders of taxpayers.

In addition, many states—even if alignment evidence via a PAD process is compelling—may still require an alignment study conducted by an independent evaluator using conventional alignment methodologies in the short term, such as Webb (2005). No matter whether alignment is engineered from the outset or a study is done post hoc, the notion of what constitutes alignment needs to broaden to one of coherence across all aspects of assessment design: learning standards, RPLDs, item design specifications, scale properties, and interpretations about what students know and can do.

Evidentiary Argument Summary

In summary, when we engage in PAD we are defining our inferential argument and collecting evidence for our evidentiary argument from the beginning of the design process, which makes creating and collecting evidence for our validation argument an ongoing endeavor, rather than a set of post hoc analyses. In PAD, evidence comes in a variety of forms and represents a broader conception of evidence than the conventional narrow definition that evidence is equivalent to the data that are collected from administered items and analyzed psychometrically. We assert that the assessment industry should hold the assessment design process to the same kinds of scrutiny that we do our psychometric analyses and other judgmental-based processes, such as standard setting.

ASSESSMENT DESIGN: LOOKING AHEAD

Assessment designers and measurement professionals in the early 21st century are extremely fortunate to have several challenging and interesting problems to solve with emergent innovations. As demand grows for more rapid feedback that supports valid inferences about students and is instructionally actionable, there is growing interest in "stealth assessment," smart games, and other mechanisms by which to leverage technology to seamlessly integrate learning and assessment in engaging ways for students (DiCerbo, 2014; Hong et al., 2019; Kim & Shute, 2015; Petrusel, 2014; Shute & Ventura, 2013; Young et al., 2011).

On a related note, at the time of this writing, the proliferation of assessment products flooding the marketplace that use one or more families of models lumped into the broad category of artificial intelligence (AI) is staggering. The models undergirding these products are developed by and large without the benefit of any assessment design

or psychometric theory and practice. The ways in which this evolution will change the field and the industry are unknown. If that were not sufficiently challenging, there is a growing desire for novel constructs to be incorporated into these blended learning and assessment systems. The scientific practices of the NGSS (e.g., developing and using models, planning and carrying out investigations) and the ACT Holistic Framework for Education and Work Success (e.g., collaborative problem-solving, sustaining effort) are examples of increasing complexity of learning and assessment needs. As the complexity of targets of measurement increases, the case for using PAD becomes stronger. It is hard to imagine a compelling validity argument without a transparent chain of inference that links all of the essential components together—for example, targets of measurement, RPLDs, items, scale properties, intended score inferences—coherently and elegantly.

Three related challenges will need to be addressed before we can realize the full potential of integrated learning and assessment and the potential of AI to revolutionize learning and assessment writ large. As we will argue, these issues cannot be addressed, much less solved, without a commitment to PAD. First is the use of accessibility features and the relationship of these features to the targets of measurement. Second is the evolution from ensuring that assessments are fair for all students and devoid of bias to assessments that reflect and value the diversity of cultures that our students represent. The third issue is the emergent focus on student motivation and engagement as it relates to estimates of their proficiency. The role of student motivation and engagement in assessment is made especially complex when considered alongside accessibility features for students with a variety of disabilities, as well as culturally and linguistically responsive assessments for students from our historically marginalized populations. These complex relationships need to be examined particularly closely for assessments that may not carry the same gravitas for students as assessments that contribute to their grades, promotion, placement, or other stakes that are important to students, such as interim assessments, assessments embedded in instruction, or other game-based or stealth assessments.

Accessibility and Cultural and Linguistic Responsiveness

It is beyond the scope of this chapter to give full treatment to the topic of accessibility, accommodations, and fairness or, as fairness is treated with a more contemporary perspective, cultural and linguistic responsiveness, in educational assessment. Other chapters in this volume are dedicated to a full treatment of these topics (see Zwick and Rodriguez & Thurlow). These issues are addressed here in light of the challenges that remain to be addressed in assessment design and development to ensure our assessments are accessible and responsive to the cultural and linguistic diversity of the students whom we serve. That said, the research and practice in these areas are developing rapidly and whatever is articulated here in 2025 is likely to be out of date in the next year or so. Nonetheless, we include examples here of how, without the precision and transparency demanded by the PAD process, incorporating accessibility and culturally and linguistically responsive features could jeopardize the validity of the inferences about what students know and

can do. As we have demonstrated in this chapter, PAD forces the detailed articulation of the assumptions about the intended targets of measurement and the intended inferences, and the resulting design specifications can be used to support decisions about accessibility and culturally and linguistically responsive features such that inferences about students from different populations are as free of interference as possible.

Let us begin with an example of an item from a standards-based, adaptive, computer-based interim assessment whose primary purpose and use is to inform instruction through categorizing students at the most beneficial point along an instructional pathway given their strengths and weaknesses in the target domain. The example item in Figure 7.17 illustrates an item designed with keen attention to both the National Center on Educational Outcomes recommendations to ensure that universal design for learning is applied to assessments and the WCAG (W3C, 2018) guidelines for accessibility features for items delivered via the Internet. These guidelines include but are not limited to simple, clear, and intuitive instructions and procedures, maximized readability and comprehensibility, and maximized legibility (National Center on Educational Outcomes, n.d.). Other features from WCAG guidelines include the following:

- strong color contrast between the shape and its background
- additional black outlining around the edges of the shape for added visual definition
- bold font used for answer selection; color contrast between black font and white background
- the mathematical expression is not an image so that an automated screen reader can read it aloud
- alternative text is included for this image so that a screen reader can be used for students who are visually challenged; for example, "A rectangular prism that is shaped like a shoebox. A legend indicates that $l = 7$, $w = 3$, and $h = 2$."
- keyboard navigation allows the student to use the keyboard arrow keys to move from element to element in the correct order; therefore, a student using a screen reader will hear the many elements in this order and can navigate forward and backward between elements as needed:
  - the stem
  - the expression
  - the alt text describing the prism and the variable values
  - the four answer choices in the order shown on screen
- Spanish transadaptation

This example of accessibility features is rather straightforward. However, the line between accessibility and infringement on the intended target of measurement can blur, for example, with the addition of an audio option for reading aloud the text, item prompt, and response options that is included for many assessment items and available to all students—even those who do not have difficulty processing written text or visual impairment. An audio option likely would not interfere with the intended target

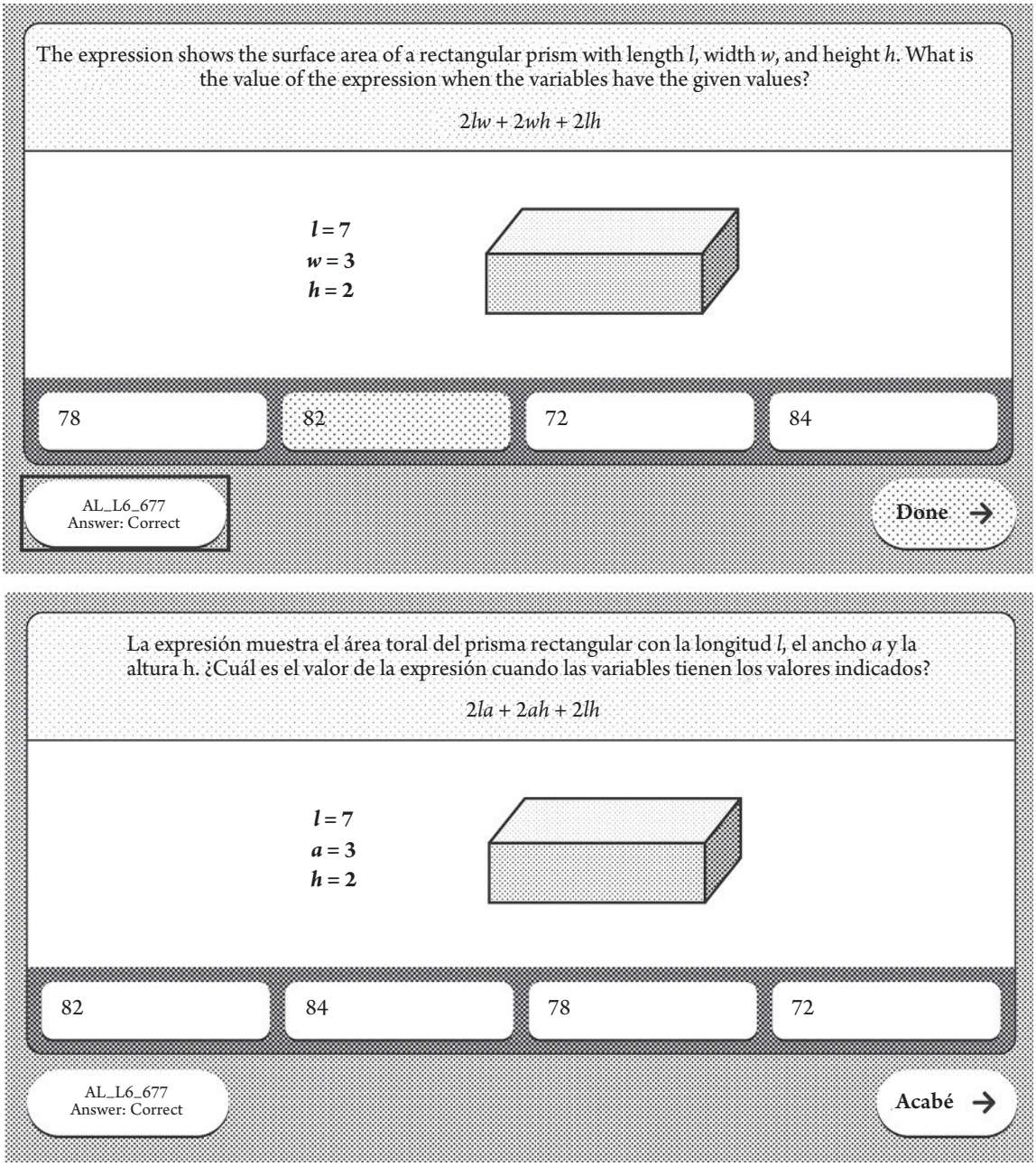

FIGURE 7.17
Example Item in English and Spanish to Illustrate Accessibility Features

of measurement for a straightforward math item like the one shown above, but would present a conflict for items that are designed to measure, for example, reading comprehension (as opposed to comprehension regardless of receptive mode). Without rigorous interrogation and documentation of the assumptions undergirding each assessment design choice, such as the kind demanded by PAD, we risk sloppy measurement and, as a result, flimsy arguments supporting our claims about what students know and can do when accessibility features are part of the item design.

Items are designed to be fair to all students (e.g., to avoid terms that privilege particular students when the concept or term is not the target of measurement) and are typically reviewed by independent panels to help ensure the items are free of sensitive topics or content (e.g., item-writing guidelines typically indicate that certain topics that could upset students be avoided, such as hurricanes or death) and bias (e.g., items do not include images or terms that portray students from marginalized communities in stereotypical or derogatory ways). At the time of this writing, our conventional notions of what constitutes fairness, sensitivity, and bias are being deeply interrogated as the assessment industry catches up with what pedagogy, curriculum, and instructional scholars and practitioners have known for quite some time: being responsive and inclusive of students' cultural and linguistic diversity matters in the learning endeavor (Gay, 2000; Hammond, 2015; Paris & Alim, 2017). In short, to be authentically culturally and linguistically responsive, we must rethink what is construct relevant and irrelevant and what is and is not sensitive to various populations (Randall, 2023; Solano-Flores, 2023). Hollie (2018) noted that Hollie (2012) defined culturally and linguistically responsive teaching as "the validation and affirmation of the home (indigenous) culture and home language for the purposes of building and bridging the student to success" (p. 23). In this context, culture has a broad definition and meaning; Hollie's work defines culture along multiple dimensions: ethnicity, sexual orientation, nationality, socioeconomic status, religion, gender, and age. For example, gardens are a favorite context for math items. Rather than having the context be a backyard (typical of suburbs) for every item that uses a garden, make sure that there is a broader spectrum of representation that includes farms (rural areas), community gardens (urban areas), and gardens from other cultures (e.g., terraced gardens from Indonesia). Another more poignant example would be to ensure that our assessment passages do not erase the lived experiences and histories of students from marginalized communities. For example, in a reading comprehension assessment where the passage topic is, for example, the 1893 World's Fair, a passage that only celebrates the wonder and achievements of the fair but fails to mention that luminaries Frederick Douglass and Ida B. Wells protested the exclusion of African Americans from exhibits (Duster, 1970) erases the history, and therefore the culture, of Black students in classrooms today. As we expand our contexts to be more responsive to and reflective of our ever-diversifying student body, we will need to

grapple with the inherent tension of attending to culturally and linguistically responsive content and contexts while simultaneously adhering to accessibility and sensitivity guidelines that steer item writers away from anything that may be potentially context rich or controversial. These debates must occur with a shared understanding of the target of measurement and what constitutes construct-relevant and construct-irrelevant variance. As these discussions become more nuanced, the precision, clarity, and transparency that are the hallmarks of PAD become more and more needed.

Engagement and Motivation

The research base on the role of student engagement and motivation in learning and assessment continues to grow. Unless students are engaged in the assessment, they will not be motivated to perform their best, and the assessment results for those students are likely underestimates of what they know and can do. This is a primary concern for many who are questioning what the lack of culturally and linguistically responsive assessments has meant for the achievement results of our students from marginalized populations (Gutiérrez, 2017; Lyiscott, 2019; Randall, 2021, 2023; Solano-Flores, 2019, 2023). For interim assessments designed primarily to inform instruction, rather than to assign grades, lack of engagement and motivation can be a real issue (Wise & DeMars, 2005). As most K–12 testing is slowly but surely moving toward computer-based rather than paper-based formats, and as the promise of blending learning and assessment to become the same endeavor unfolds, assessment design is in need of partnership with user experience designers (UX designers, UX design). User experience, in this usage, encompasses the various elements of design expertise that go into creating effective items, including interaction design, user interface design, art creation and curation, and usability.

Experts in UX design bring a perspective to assessment design that is generally not represented in interdisciplinary teams of psychometricians, content experts, educators, item writers, and learning scientists. Conventional practice would rely on illustrators or photo editors to participate in a very narrow way, such as providing a graphic as part of an item stimulus after an art specification has been defined by the item writer, which is a questionable practice given that said art specifications could benefit greatly from trained UX designers. However, UX designers, when part of the team from the start of the assessment design process, can help us think through the student experience of the assessment in new and compelling ways. The basic goals of UX design are to create an experience that identifies and meets the needs of the user. Some of the hallmarks of a well-designed experience are simplicity, transparency of goals and actions, clear communication of information, and an experience that is enjoyable to use. These goals translate to the experience of student assessment as well. There are more points of entry for UX design to partner in a computer-based test than a paper-based test, but UX designers also have perspective and expertise to offer in paper-based testing, since the goals and principles of good communication design apply to all media. In computer-based testing, UX designers are concerned

with not only the graphics required for items, but also the student navigation experience—from starting the assessment, to moving between items, to the within-item experience. The goal is to ensure that everything presented to the student and each part of the navigation process is intuitive and does not get in the way of the intended purpose: to optimally measure what the student knows and can do. In other words, the UX designer is just as committed to avoiding construct-irrelevant variance as the item writers and psychometricians. UX designers help assessment designers avoid construct-irrelevant variance through probing the assumptions that undergird item design with questions like:

- Are the interactions clear and easy to use? Are they age appropriate?
- Is the art content accessible and equitable? Does it represent the word, object, or concept in an unambiguous way the student can understand?
- Are there superfluous, decorative, or distracting elements within an artwork that may mislead or inhibit a student?
- Does the experience feel familiar and consistent from item to item?
- Are items designed in such a way that stimulus, stem, distractors, and answer areas are consistent across item types so that a student needs little time to figure out how to input their answer?
- Does the UX support taking the assessment over one session or multiple sessions?
- Does the user have appropriate context for what they are doing? Do they know why they are doing it?
- Does the hierarchy of content and images within an item inform the student as to where and how to answer the question?

With a PAD approach to assessment design, these types of questions would be considered in the development of the design patterns and the assessment delivery model. A strong partnership with the UX design team will help make these models better.

CHAPTER SUMMARY

Industry Inflection Point for PAD

What might be the future for the use of PAD? Andy Grove, Intel's former chief executive officer, observed, "When spring comes, snow melts first at the periphery, because that is where it is most exposed" (McGrath, 2019, p. 14). Might the same be true of the assessment industry? We might find the future influences on PAD not in the large-scale admissions tests or the state summative assessments, but in the game-based and stealth assessments, the assessment of emerging constructs like three-dimensional science learning, and the assessment of noncognitive constructs such as social-emotional learning.

PAD has been available to the field of educational assessment for over 20 years (Mislevy et al., 2002). Prior to the work of Mislevy and colleagues, the foundational concepts

for PAD were described by other scholars working at the intersection of learning, cognition, and assessment, including but not limited to Lohman and Ippel (1993), Nichols (1994), Embretson (1998), Pellegrino et al. (1999), and Snow and Lohman (1989). The use of PAD appears to offer many benefits to assessment developers and users. Yet, PAD does not enjoy widespread use in operational educational assessment design and development. Pieces have been implemented here and there in organizations and testing programs. For example, a version of PAD (ECD) was used in the design of the National Board for Professional Teaching Standards (Pearlman, 2008a) and the Test of English as a Foreign Language (Pearlman, 2008b). It was also used in the redesign of the AP program (Huff & Plake, 2010a) and by the Cisco learning network (Behrens et al., 2010), and it was required by the 2009 Race to the Top legislation that supported the assessment consortia (U.S. Department of Education, 2010). PAD is also being used by organizations to design assessments to measure the NGSS (Harris et al., 2019; Luecht, 2019). However, at the time of this writing, the authors are unaware of any operational large-scale assessment programs that systemically implement all the essential elements of PAD. The lack of widespread PAD adoption suggests that funders and users currently perceive less value in PAD compared to conventional test development.

The popularity of PAD may be waiting for an assessment industry strategic inflection point to dramatically increase PAD's value to assessment developers and users. A strategic inflection point is a change in the business environment that throws some assumptions into question and upends the basic assumptions of a business model. When this happens, we posit that the perceived value of PAD will outweigh the perceived value of conventional assessment development. Conventional assessment development and psychometric practices yield a sufficient return on investment for test developers. What is needed is a large-scale counterexample that demonstrates that PAD maximizes return on investment in assessment development.

Leading indicators—things that are not yet undisputed facts in an industry—suggest that the assessment industry is creeping toward an inflection point for PAD. A leading indicator for this inflection point is that the constructs of interest are becoming more complex and less approachable using conventional assessment development (Nichols & Huff, 2017). In addition, educators and policy makers in states and districts are increasingly unhappy with both the amount of testing and the quality of educational assessments. As such, we expect educational assessments to simultaneously meet multiple purposes and uses and to be more transparent with regard to their quality and meaningfulness in teaching and learning. Finally, assessment designers are under pressure, and rightfully so, to make sure our assessments are accessible to all students, are culturally and linguistically responsive, and are engaging to boot. The testing industry has the power of technology on its side in meeting these challenges, especially as it becomes easier for interdisciplinary teams to collaborate remotely. In addition, learning science research is occurring at a rapid rate that helps us design assessments that better support inferences about what students know and can do. Taken together, maybe there is an inflection point on the near horizon.

The arrival of inflection points is difficult to predict. For example, Reed Hastings, the founder of Netflix, waited and waited for the inflection point for the online streaming model. As Hastings explained,

> In 1997, we said that 50% of the business would be from streaming by 2002. It was zero. In 2002, we said that 50% of the business would be from streaming by 2007. It was zero. . . . Now streaming has exploded. . . . We were waiting for all these years. Then we were in the right place at the right time. (McGrath, 2019, p. 79)

Similarly, Mislevy et al. (2002) were anticipating an inflection point in the testing industry over 20 years ago:

> Standard procedures for designing and carrying out assessments have worked satisfactorily for the assessments we have all become familiar with over the past half century. Their limits are sorely tested today. The field faces demand for more complex inferences about students, concerning finer grained and interrelated aspects of knowledge and conditions under which this knowledge can be to bear. Advances in technology can provide far richer samples of performances, in increasingly realistic and interactive settings; how can we make sense of this complex data? And even with familiar assessments, cost pressures from continuous testing and social pressures for validity arguments demand more principled assessment designs and operations. (p. 126)

The philosophical orientation, practices, and procedures of PAD must be fully embraced and operationalized by each organization and each individual within who is responsible for designing and developing assessments of educational achievement, whether blended with instruction, interim, or summative. Our students, parents, teachers, principals, district staff, state educational leaders, and the public whom we serve deserve our best efforts to support the educational endeavor. Conventional approaches to educational assessment do not represent our best thinking and our best work; we can do better and we must.

ACKNOWLEDGMENTS

The authors would like to thank the editors and reviewers for their thoughtful comments and questions that made this chapter—and our thinking—crisper and clearer. Special gratitude goes to our mentors and pioneers in assessment design Mari Pearlman, Linda Steinberg, Ric Luecht, Steve Ferrara, and Bob Mislevy. We are grateful to Eli Mintzer for formatting the figures.

REFERENCES

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

Anastasi, A. (1967). Psychology, psychologists, and psychological testing. *American Psychologist, 22*(4), 297–306.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1–15.

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

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). *An evidence centered design for learning and assessment in the digital world* (CRESST Report 778). University of California, National Center for Research on Evaluation, Standards, and Student Testing. https://files.eric.ed.gov/fulltext/ED520431.pdf

Bejar, I. I. (2010). Application of evidence-centered assessment design to the Advanced Placement redesign: A graphic restatement. *Applied Measurement in Education, 23*(4), 378–391. https://doi.org/10.1080/08957347.2010.510969

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). JAM Press.

Bennett, R. E. (2015). The changing nature of educational assessment review of research in education. *Review of Research in Education, 39*(1), 370–407.

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

Brookhart, S. M. (2009). *Exploring formative assessment*. Association for Supervision and Curriculum Development.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32–42.

CAST. (2018). *Universal Design for Learning Guidelines*, version 2.2. http://udlguidelines.cast.org

College Board. (2019). *AP [course]: Course and exam description*. https://aphighered.collegeboard.org/courses-exams/course-exam-redesign

Council of Chief State School Officers & Association of Test Publishers. (2013). Operational best practices for statewide large-scale assessment programs. https://ccsso.org/sites/default/files/2023-09/ATP%20Best%20Practices%20Version%202-FINAL-082113-interior-300dpi.pdf

DiCerbo, K. E. (2014). Game-based assessment of persistence. *Journal of Educational Technology & Society, 17*(1), 17–28.

Duster, A. M. (Ed.). (1970). *Crusade for justice: The autobiography of Ida B. Wells.* The University of Chicago Press.

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

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

Ercikan, K., & Seixas, P. (Eds.). (2015). *New directions in assessing historical thinking*. Routledge.

Ewing, M., Packman, S., Hamen, C., & Thurber, A. C. (2010). Representing targets of measurement within evidence-centered design. *Applied Measurement in Education, 23*(4), 325–341.

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). National Council on Measurement in Education and American Council on Education.

Ferrara, S., Dogan, E., Glazer, N., Gorin, J., Haberstroh, J., Hain, B., Huff, K., Larkin, J., Nichols, P., Piper, C., & Sheehan, P. (2014, April 3–7). Development of cognitive complexity measures for PARCC (Presentation). In A. Rupp (Chair), *Cognition and assessment SIG business meeting and poster session.* American Educational Research Association Annual Meeting, Philadelphia, PA, United States.

Ferrara, S., & Steedle, J. (2015, April 15–19). *Predicting item parameters using regression trees: Analyzing existing item data to understand and improve item writing* [Paper presentation]. National Council of Measurement in Education Annual Meeting, Chicago, IL, United States.

Ferrara, S., Steedle, J., & Kinsman, A. (2015). *PARCC cognitive complexity: Analysis 1, 2, and 3 results*. Partnership for Assessment of Readiness in College and Careers. https://eric.ed.gov/?id=ED599050

Frederiksen, J. R., & Collins, A. (1998). A systems approach to educational testing. *Educational Researcher, 18*(9), 27–32.

Gay, G. (2000). Culturally responsive teaching: Theory, research, and practice. Teachers College Press.

Gillmor, S. C., Poggio, J., & Embretson, S. (2015). Effects of reducing the cognitive load of mathematics test items on student performance. *Numeracy, 8*(1). https://doi.org/10.5038/1936-4660.8.1.4

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*(9), 923–936.

Goldstein, D. (2020, February 15). An old and contested solution to boost reading scores: Phonics. *New York Times*. https://www.nytimes.com/2020/02/15/us/reading-phonics.html

Gutiérrez, R. (2017). Political conocimiento for teaching mathematics: Why teachers need it and how to develop it. In S. Kastberg, A. M. Tyminski, A. Lischka, & W. Sanchez (Eds.), *Building support for scholarly practices in mathematics methods* (pp. 11–38). Information Age Publishing.

Hammond, Z. L. (2015). *Culturally responsive teaching and the brain*. Corwin Press.

Hanford, E. (2018, October 26). Why are we still teaching reading the wrong way? *New York Times*. https://www.nytimes.com/2018/10/26/opinion/sunday/phonics-teaching-reading-wrong-way.html

Hanford, E. (2019, December 5). There is a right way to teach reading, and Mississippi knows it. *New York Times*. https://www.nytimes.com/2019/12/05/opinion/mississippi-schools-naep.html

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice, 38*(2), 53–67.

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & McElhaney, K. W. (2016). *Constructing assessment tasks that blend disciplinary core ideas, crosscutting concepts, and science practices for classroom formative applications*. SRI International. https://cadrek12.org/sites/default/files/Harris_Krajcik_Pellegrino_McElhaney_constructing%20assessment%20tasks%202016.pdf

Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., & Spurgeon, L. (2015). *Student testing in America's great city schools: An inventory and preliminary analysis*. Council of the Great City Schools. http://www.cgcs.org/cms/lib/DC00001581/Centricity/Domain/87/Testing%20Report.pdf

Hendrickson, A., Ewing, M., Kaliski, P., & Huff, K. (2013). Evidence-centered design: Recommendations for implementation and practice. *Journal of Applied Testing Technology, 14*(14).

Hendrickson, A., Huff, K., & Luecht, R. (2010). Claims, evidence, and achievement level descriptors as a foundation for item design and test specifications. *Applied Measurement in Education, 23*, 358–377.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative classroom assessment? *Educational Measurement: Issues and Practice, 28*(3), 24–31.

Herman, J. L. (2010). *Coherence*: *Assessment success key to next generation* [AACC report]. University of California. https://files.eric.ed.gov/fulltext/ED524221.pdf

Hollie, S. (2012). *Culturally and linguistically responsive teaching and learning: Classroom practices for student success* (1st ed.). Shell Education.

Hollie, S. (2018). *Culturally and linguistically responsive teaching and learning: Classroom practices for student success* (2nd ed.). Shell Education.

Hong, J. C., Chang, C. H., Tsai, C. R., & Tai, K. H. (2019). How situational interest affects individual interest in a STEAM competition. *International Journal of Science Education, 41*(12), 1667–1681.

Hong, J., & Lissitz, R. W. (Eds.). (2017). *Test fairness in the new generation of large-scale assessment*. Information Age Publishing.

Huff, K., Alves, C. B., Pellegrino, J., & Kaliski, P. (2013). Using evidence-centered design task models in automatic item generation. In M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 102–113). Routledge.

Huff, K., & Ferrara, S. (2010, June). *Frameworks for considering item response demands and item difficulty* [Paper presentation]. Council of Chief State School Officers National Conference on Large-Scale Assessment, Detroit, MI, United States.

Huff, K., & Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 19–60). Cambridge University Press.

Huff, K., & Plake, B. (2010a). Evidence-centered assessment design in practice. *Applied Measurement in Education, 23*, 307–309.

Huff, K., & Plake, B. (2010b). Innovations in setting performance standards for K–12 test-based accountability. *Measurement: Interdisciplinary Research and Perspectives, 8*(2), 130–144.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*, 310–324.

Huff, K., Warner, Z., & Schweid, J. (2017). Large-scale standards-based assessments of educational achievement. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 397–426). Wiley–Blackwell.

Hurford, D. P., Hurford, J. D., Head, K. L., Keiper, M. M., Nitcher, S. P., & Renner, L. P. (2016). The dyslexia dilemma: A history of ignorance, complacency and resistance in colleges of education. *Journal of Childhood & Developmental Disorders, 2*(3). doi: 10.4172/2472-1786.100034

Improving America's Schools Act of 1994. Pub. L. No. 103–382 (1994).

Johannesson, P., & Perjons, E. (2014). *An introduction to design science*. Springer International Publishing.

Johnstone, C., Altman, J., Thurlow, M., & Moore, M. (2006). *Universal design online manual*. University of Minnesota, National Center on Educational Outcomes. https://nceo.info/Resources/publications/UDmanual/UDmanualPrint.htm

Kaliski, P., Huff, K., & Barry, C. (2011, April). *Aligning items and achievement levels: A study comparing expert judgments* [Paper presentation]. Meeting of the National Council on Measurement in Education, New Orleans, LA, United States.

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

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

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

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in game-based assessment. *Computers & Education, 87*, 340–356.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science–practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798.

Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction*. Lawrence Erlbaum Associates.

Lane, S., Raymond, M., & Haladyna, T. (Eds). (2016). *Handbook of test development* (2nd ed., pp. 119–143). Routledge.

Lawrence, I. M., & Shea, E. C. (2008). *Improving assessment: The intersection of psychology and psychometrics* (ETS Research Memorandum No. RM-08-15). ETS.

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*, 3–16.

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.

Lewis, D., & Cook, R. (2020). Embedded standard setting: Aligning standard-setting methodology with contemporary assessment design principles. *Educational Measurement: Issues and Practice, 39*(1), 8–21.

Lindquist, E. F. (Ed.). (1951). *Educational measurement*. American Council on Education.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694.

Lohman, D. F., & Ippel, M. J. (1993). *Cognitive diagnosis: From statistically based assessment toward theory-based assessment.* In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 41–71). Lawrence Erlbaum Associates.

Louisiana Department of Education. (2020). *A teacher's guide to LEAP 360*. https://www.louisianabelieves.com/docs/default-source/assessment/a-teachers-guide-to-leap-360.pdf?sfvrsn=4

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology, 14*(1), 1–38.

Luecht, R. M. (2019, January). *Strengthening Claims-Based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS): The role of performance level descriptors for establishing meaningful and useful reporting scales in a principled design approach* [White paper]. Nebraska Department of Education. https://www.scillsspartners.org/wp-content/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf

Lyiscott, J. (2019). *Black appetite. White food: Issues of race, voice, and justice within and beyond the classroom*. Routledge.

Marion, S., Thompson, J., Evans, C., Martineau, J., & Dadey, N. (2019). *The challenges and opportunities of balanced systems of assessment: A policy brief*. National Center for the Improvement of Educational Assessment. https://files.eric.ed.gov/fulltext/ED598421.pdf

McGrath, R. (2019). *Seeing around corners: How to spot inflection points in business before they happen*. Houghton Mifflin Harcourt.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education.

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

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). American Council on Education.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483.

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

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

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology* (PADI Technical Report 9). SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). On the roles of task model variables in assessment design. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 97–128). Lawrence Erlbaum Associates.

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 Assessment Governing Board. (2018). *Technology & engineering literacy framework for the 2018 National Assessment of Educational Progress*. https://www.nagb.gov/content/nagb/assets/documents/publications/frameworks/technology/2018-technology-framework.pdf

National Assessment Governing Board. (2021a). *Mathematics framework for the 2026 National Assessment of Educational Progress*. https://www.nagb.gov/naep-subject-areas/mathematics/2026-naep-mathematics-framework.html

National Assessment Governing Board. (2021b). *Reading framework for the 2026 National Assessment of Educational Progress*. https://www.nagb.gov/naep-subject-areas/reading/framework-archive/2026-reading-framework.html

National Center on Educational Outcomes. (n.d.). *Universal design of assessments: Overview*. Retrieved September 14, 2020, from https://nceo.info/Assessments/universal_design/overview

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. http://www.corestandards.org/read-the-standards/

National Research Council. (2000). *How people learn: Brain, mind, experience, and school*. The National Academies Press. https://doi.org/10.17226/6160

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. The National Academies Press. https://doi.org/10.17226/10019

National Research Council. (2002). *Learning and understanding: Improving advanced study of mathematics and science in U.S. high schools*. The National Academies Press. https://doi.org/10.17226/10129

National Research Council. (2005). *How students learn: History, mathematics, and science in the classroom*. The National Academies Press. https://doi.org/10.17226/10126

NGSS Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academic Press. https://www.nextgenscience.org/search-standards

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

Nichols, P., Ferrara, S., & Lai, E. (2016). Principled design for efficacy: Design and development for the next generation of assessments. In H. Jiao & R. W. Lissitz (Eds.), *The next generation of testing: Common Core Standards, Smarter-Balanced, PARCC, and the nationwide testing movement* (pp. 49–81). Information Age Publishing.

Nichols, P., & Huff, K. (2017). Assessments of complex thinking. In K. Ercikan & J. Pellegrino (Eds.), *Validation of score meaning in the next generation of assessments: The use of response processes* (pp. 63–74). Routledge.

Nichols, P., Kobrin, J. L., Lai, E., & Koepfler, J. D. (2017). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications, first edition* (pp. 13–40). Wiley–Blackwell.

Nichols, P., Meyers, J., & Burling, K. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice, 28*(3), 14–23.

No Child Left Behind Act of 2002. Pub. L. No. 107–110, § 115, Stat.1425 (2002).

Paris, D., & Alim, H. S. (2017). *Culturally sustaining pedagogies: Teaching and learning for justice in a changing world*. Teacher's College Press.

Pearlman, M. (2008a). Chapter 3: The design architecture of NBPTS certification assessments. In R. E. Stake, S. Kushner, L. Ingvarson, & J. Hattie (Eds.), *Assessing teachers for professional certification: The first decade of the National Board for Professional Teaching Standards: Advances in program evaluation* (Vol. 11, pp. 55–91). Emerald Group Publishing.

Pearlman, M. (2008b). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). Routledge.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Research in Education, 24*, 307–352.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2011). *Knowing what students know: The science and design of educational assessment*. National Academies Press. https://www.nap.edu/catalog/10019/knowing-what-students-know-the-science-and-design-of-educational

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments, *Educational Psychologist, 51*(1), 59–81.

Penuel, W. R., & Shepard, L. A. (2017). Social models of learning and assessment. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 146–173). Wiley–Blackwell.

Perie, M. (2006). *Convening an articulation panel after a standard setting meeting: A how-to guide*. Center for Assessment. https://www.nciea.org/publications/RecommendforArticulation_MAP06.pdf

Perie, M., & Huff., K. (2016). Determining the content and cognitive demand for achievement tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 119–143). Routledge.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice, 28*(3), 5–13. https://doi.org/10.1111/j.1745-3992.2009.00149.x

Petrusel, R. (2014). Integrating click-through and eye-tracking logs for decision-making process mining. *Informatica Economica, 18*(1), 56.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82–90.

Randall, J. (2023). It ain't near 'bout fair: Re-envisioning the bias and sensitivity review process from a justice-oriented antiracist perspective. *Educational Assessment, 28*(2), 68–82.

Ruiz-Primo, M. A., Furtak, E. M., Ayala, C. C., Yin, Y., & Shavelson, R. J. (2010). Formative classroom assessment, motivation, and science learning. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative classroom assessment* (pp. 139–158). Routledge.

Rupp, A. A., & Leighton, J. P. (Eds). (2017). *The handbook of cognition and assessment: Frameworks, methodologies, and applications*. Wiley–Blackwell.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). National Council on Measurement in Education and American Council on Education.

Schneider, M. C. (2017, April 26–30). *Using principled assessment design to support formative assessment and students' opportunities to learn* [Paper presentation]. National Council on Measurement in Education Annual Meeting, San Antonio, TX, United States.

Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action. *Applied Measurement in Education, 26*(3),159–162.

Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education, 26*(3), 191–204.

Schneider, M. C., Huff, K., Egan, K., Gaines, M., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement level descriptors. *Educational Assessment, 18*(2), 99–121.

Schneider, M. C., & Johnson, R. L. (2018). *Using formative assessment to support student learning objectives*. Routledge.

Schneider, M. C., & Meyer, J. P. (2012). Investigating the efficacy of a professional development program in formative classroom assessment in middle school English language arts and mathematics. *Journal of Multidisciplinary Evaluation, 8*(17), 1–24.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Wiley.

Schum, D. A. (2009). Science of evidence: Contributions from law and probability. *Law, Probability, and Risk, 8*, 197–231.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32–37.

Shute, V. J., & Ventura, M. (2013). *Measuring and supporting learning in games: Stealth assessment*. The MIT Press.

Simon, H. A. (1969). *The sciences of the artificial*. Massachusetts Institute of Technology.

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). American Council on Education/Macmillan.

Solano-Flores, G. (2019). Examining cultural responsiveness in large-scale assessment: The matrix of evidence for validity argumentation. *Frontiers in Education, 4*, 43. https://doi.org/10.3389/feduc.2019.00043

Solano-Flores, G. (2023). How serious are we about fairness in testing and how far are we willing to go? A response to Randall and Bennett with reflections about the *Standards for Educational and Psychological Testing*. *Educational Assessment, 28*(2), 105–117.

South Carolina Department of Education. (n.d.). *Adoption list of formative assessments*. https://ed.sc.gov/tests/middle/adoption-list-of-formative-assessments/

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S. E. (2006). Reasoning in theory and practice. In D. L. Hitchcock & B. Verheij (Eds.), *Arguing on the Toulmin model. New essays in argument analysis and evaluation* (pp. 25–29). Springer.

U.S. Department of Education. (2010). *Race to the Top program guidance and frequently asked questions*. Authorized under Sections 14005 and 14006 of the American Recovery and Reinvestment Act of 2009. https://www2.ed.gov/programs/racetothetop/faq.pdf

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal, 115*(2), 270–289.

van Aken, J. E., & Romme, G. (2009). Reinventing the future: Adding design science to the repertoire of organization and management studies. *Organization Management Journal, 6*(1), 5–12.

Webb, N. L. (2005). *Web Alignment Tool (WAT): Training manual* (Draft Version 1.1). Wisconsin Center for Education Research, Council of Chief State School Officers.

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

Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education, 33*(3), 1–12.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*, 1–17.

World Wide Web Consortium. (2018). *Web Content Accessibility Guidelines (WCAG) 2.1*. https://www.w3.org/TR/WCAG21/

Young, V. M., House, A., Wang, H., Singleton, C., & Klopfenstein, K. (2011, May). *Inclusive STEM schools: Early promise in Texas and unanswered questions* [Paper presentation]. Highly successful STEM schools or programs for K–12 STEM education: A workshop, Washington, DC, United States. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_072639.pdf

NOTES

1. We believe that classroom assessments could also benefit from principled assessment design principles and practices, but that is beyond the scope of this chapter.
2. For the purposes of this chapter, we are addressing students who take generalized assessments. Alternate assessments for students with the most significant cognitive disabilities are discussed elsewhere in this volume, including by Lane and Marion (in relation to validity) and Rodriguez and Thurlow (in relation to fairness).
3. For the purposes of this chapter, *task* and *item* are used interchangeably.
4. This chapter uses the term *item writer* to refer to all contributors to item development; however, in practice item writer generally refers specifically to those who write initial drafts of items. Most of the work attributed to item writers in this chapter is performed by content development professionals who shepherd those initial item drafts through the full item development process, which includes an iterative cycle of review and editing. These educational measurement professionals often have titles like *content specialist* or *editor.*
5. "Off the shelf" is an industry term that refers to tests developed for use in any state, regardless of the learning standards in the state.