CASE STUDY

Designing and Developing Educational Assessments for Contemporary Needs

Kristen Huff

Designing and Developing Educational Assessments for Contemporary Needs

Kristen Huff

Abstract

Conventional assessment design and development approaches that have served the field for decades are struggling to meet today's growing educational needs. To adequately handle the complexity and scope of the knowledge we aim to measure, assessments must be designed with as much rigor and clarity as possible. The Principled Assessment Design approach provides solutions to these challenges that remain coherent across all elements of an assessment system.

The three iterative phases of PAD foster a deeper shared understanding of student cognitive processes and the role of performance level descriptors. Rather than reactive, ad hoc validation based on analysis of empirical data, a strong inferential, evidentiary, and validation argument begins to take shape proactively in the design process. The documentation generated throughout each iteration provides design tools that inform development of performance level descriptors (PLDs) and task models.

PAD therefore offers an adaptable framework to address evolving educational goals while embracing our growing understanding of responsiveness to learner strengths and needs. The principles of PAD are essential in creating fair and accessible assessments while maintaining the integrity of the constructs being measured. The assessment community faces an inflection point amid complex and contemporary demands, expectations, and capabilities; proponents of PAD are poised to meet those needs.

Introduction

Educational assessment stands at a critical inflection point. Conventional assessment design and development approaches that have served the field for decades are being challenged by increasing demands for assessments that are more authentic, informative, actionable, engaging, and accessible for all students. Principled Assessment Design (PAD) is an alternative approach that addresses the challenge of creating culturally and linguistically responsive, yet fair and well-designed, assessments. Using the three iterative phases of PAD assessment designers can address these goals from the outset with strong inferential, evidential, and validation arguments. Through a focus on student cognition, consistent documentation, and continuous re-evaluation, this knowledge base can be built upon to evolve with our understanding of fairness in assessment.

The Case for Principled Assessment Design

Historically, large-scale assessments were designed primarily to rank order students, and the primary, and perhaps only, interpretation from the resulting scores was the percentile rank of the test taker in comparison to a national norm-referenced distribution (National Research Council, 2001; Shepard, 2000). The objective was not to support claims about what a student knows and can do in the tested domains. Although many assessments are now called upon to provide strong inferences about what students have learned and what they need to learn to reach a particular performance level, assessment design and development practices are still largely rooted in the norm-referenced paradigm.

There are at least three reasons that conventional approaches to assessment design and development are insufficient for meeting today's educational needs. First, there is a growing complexity of what we aim to measure, such as mathematical practices, three-dimensional science learning, and collaborative problem-solving (NGA Center & CCSSO, 2010; NGSS Lead States, 2013). Second, users are demanding that K–12 assessments serve multiple purposes. Rather than add to the proliferation of testing that occurs within any given school year, it is incumbent upon the industry to design tests from the outset that have clear, strong validation arguments that can be built upon for additional use cases (AERA, APA, & NCME, 2014; Hart et al., 2015; Huff & Goodman, 2007). Third, assessment quality is under more scrutiny than ever (OESE, 2018). Our industry needs to do a better job at ensuring that assessments that are designed to make claims about student learning are designed with as much rigor and clarity as possible to do just that.

Historically, large-scale assessments were designed to rank students—not to reveal what they know and can do.

Principled Assessment Design (PAD) is an approach to assessment design that offers solutions to these contemporary challenges. PAD is a set of practices and documentation that helps ensure coherence across all elements of an assessment system. The provenance of PAD is rich, including but not limited to construct-centered measurement (Messick, 1994; Wilson, 2005), cognitive design systems (Embretson, 1998; Rupp & Leighton, 2017), evidence-centered design (Huff et al., 2010; Mislevy, 2006; Mislevy et al., 2003; Mislevy & Haertel, 2006; Pearlman 2008a, 2008b), principled design for efficacy (Nichols et al., 2016) and assessment engineering (Luecht, 2013).

There are at least three characteristics of PAD that distinguish it from conventional assessment design. The first is the role of learning science (Pellegrino et al., 2016). In PAD, where the primary purpose of the assessment is to measure where students are along a learning trajectory, the construct must be defined and all decisions about assessment design must be rooted in the science of how students learn and build knowledge. The second is the practice of documentation that can serve as design tools throughout assessment development and assessment interpretation (and if needed, even beyond, to curricular, instructional, or teacher professional learning materials). The third is a mindset of inquiry, where reasoning from imperfect evidence requires continuous interrogation of assumptions about the inferential, evidential, and validation arguments.

There are three essential phases in PAD. These phases are iterative rather than linear. The first is analyzing the domain with respect to the science of how students learn and build knowledge (Ewing et al., 2010; Pellegrino et al., 2016). In this phase, the construct and the targets of measurement for the assessment are defined. The next phase is modeling the domain, where the approach to cognition for the assessment is defined, as well as the performance level descriptors (PLDs). The final phase of design is to create design patterns for the tasks (items) which requires a shared understanding of the difficulty drivers for the tasks (Luecht, 2019). The difficulty drivers are directly informed by the approach to cognition that is defined in the second phase.

Just as there is iteration between the three general phases of PAD, there is not necessarily a clear distinction of when design ends and development begins. The point is that there needs to be a design phase to develop a shared understanding of all of the above—and the appropriate documentation—rather than simply articulating the test specifications and starting item development, which is common practice in conventional test development (Schmeiser & Welch, 2006).

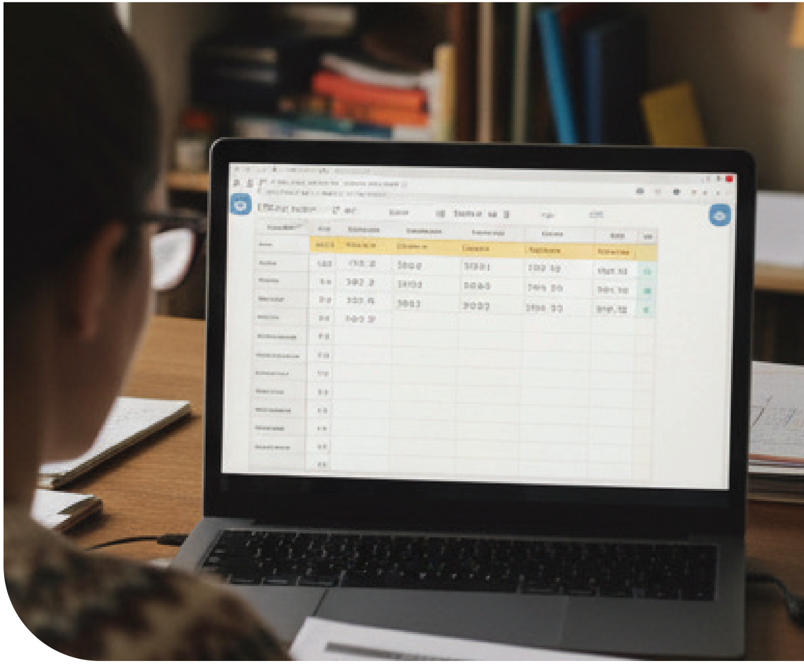
The Cognitive Approach in PAD

As mentioned above, one of the distinguishing characteristics of PAD is the explicit grounding in learning science, and as a result, a deeper focus on the role of cognition in both learning and the assessment of learning. A shared understanding—and clear documentation—of how student cognitive processes will be treated in PLDs and therefore in the assessment is a key component of the PAD.

When defining the cognitive approach that will be used to develop PLDs and therefore tasks, the interdisciplinary assessment design team investigates the answers to questions like the following, develops a shared understanding of the answers, and documents the answers in a way that can be used as design tools in the development of PLDs and, later, task models:

1. What does learning science say about how student thinking evolves from novice to proficient in this domain?
2. What constitutes observable evidence of that thinking?
3. How rigorous should proficiency be for this grade level?
4. How much learning should be expected in a given grade level for this domain?
5. What skills should be taught and assessed in this grade level to support the student learning journey?
6. What is observable evidence of each skill at each novice, proficient, advanced for this grade level?

Engaging in this line of inquiry among the assessment design team has many benefits, not the least of which is a strong inferential, evidentiary, and validation argument that is defined starting with design rather than the way validation typically occurs in conventional assessment design: post-hoc and based not solely but mostly on analyses of empirical data.



Performance Level Descriptors: The Backbone of Assessments Designed to Measure Learning

PLDs play a critical role in ensuring that the assessment is designed to support inferences about where students are along a latent proficiency continuum. In conventional approaches to assessment development, PLDs were developed after the assessment was developed as an input into the standard-setting process. In PAD, PLDs embody the approach to cognition and are the foundation of task design, informing the desired psychometric properties of the scale, and score interpretation.

Addressing Key Challenges in Assessment Design

Accessibility and Cultural Responsiveness

Assessment designers face the dual challenge of creating assessments that are accessible to all students while maintaining the integrity of the constructs being measured. For example, audio options that read text aloud may be essential accommodations for some students but could fundamentally alter what's being measured in a reading comprehension assessment. In these cases, we argue that assessment designers must have clear, strong inferential, evidential, and validation arguments so that they can nimbly engage in discussions about what inferences about student learning can and cannot be supported with various accessibility features, and whether a redefinition of the target of measurement is warranted.

Similarly, our understanding of fairness in assessment is evolving beyond merely avoiding bias to actively embracing cultural and linguistic responsiveness. Rather than simply stripping items of cultural context, there is growing recognition that assessments should reflect and value the diversity of students' lived experiences and cultural backgrounds. This might mean including garden contexts that represent urban community gardens and rural farms rather than exclusively suburban backyards, or ensuring that historical passages don't erase the experiences of marginalized communities.

PAD provides the structured framework needed to thoughtfully navigate these tensions. By clearly articulating the intended targets of measurement and assumptions about what constitutes construct-relevant versus construct-irrelevant variance, assessment designers can make principled decisions about accessibility features, accommodations, and cultural representation (CAST, 2018; Solano-Flores, 2019; World Wide Web Consortium, 2018).

Student Engagement and Motivation

The role of student engagement and motivation in assessment performance has gained increased attention, particularly for assessments that lack direct consequences for students, such as interim or embedded assessments. If students aren't engaged nor motivated to perform at their best, assessment results likely underestimate their learning (Tsai et al., 2020; Wise & DeMars 2005).

This challenge highlights the importance of integrating User Experience (UX) design expertise into assessment development. UX designers bring crucial perspectives on creating assessment experiences that are intuitive, transparent, and even enjoyable. They help ensure that navigation systems, visual elements, and interactions don't introduce construct-irrelevant barriers to performance.

Key questions that UX designers help assessment teams address include:

- Are interactions clear, easy to use, and age-appropriate?
- Is visual content accessible, equitable, and unambiguous?
- Does the experience feel familiar and consistent across items?
- Are there distracting elements that might interfere with performance?

Within a PAD framework, UX considerations become integral to task design rather than superficial enhancements. When design patterns explicitly address engagement factors and cognitive load considerations, the resulting assessments are more likely to elicit compelling evidence of where students are along their learning journey.

Looking Ahead: The Inflection Point for Assessment Design

The assessment field may be approaching what business strategists call a "strategic inflection point"—a moment when fundamental assumptions are challenged and business models are upended (Christensen, 1997; Grove, 1996). Several indicators suggest this inflection point may be imminent:

- The constructs we seek to measure are becoming increasingly complex
- Educators and policymakers are dissatisfied with both the quantity and quality of current assessments
- There are growing demands for assessments to serve multiple purposes simultaneously
- Technology is creating new possibilities for assessment design and delivery—especially AI
- Learning science research is advancing at a rapid pace

The assessment field may be approaching a strategic inflection point—a moment when fundamental assumptions are upended.

As Reed Hastings of Netflix discovered with streaming video, the timing of inflection points is difficult to predict (Hastings & Meyer, 2020). However, assessment designers who embrace principled approaches now will be well-positioned when the industry reaches its tipping point.

The adoption of PAD, with its emphasis on cognitive foundations and explicit design rationales, represents a significant shift from conventional assessment development. Like Hastings waiting for streaming to take off, proponents of PAD have been anticipating its widespread adoption for over two decades. The convergence of complex measurement demands, technological capabilities, and evolving stakeholder expectations may finally create the conditions for PAD to become standard practice rather than the exception.

Conclusion

The future of educational assessment lies in the thoughtful integration of principled design approaches, cultural responsiveness, engaging user experiences, and emerging technologies. By building assessments that provide insights, deliver meaningful information, and cohere with instruction, we can fulfill the promise of assessment as a tool that genuinely supports teaching and learning.

The assessment community faces a choice: continue with conventional approaches that have served adequately in the past or embrace more rigorous, transparent methods that can meet complex, contemporary demands. PAD, with its explicit cognitive models and carefully constructed PLDs, offers a framework not just for better assessments, but for educational experiences that truly reveal what students know and can do—and point the way toward their continued growth.

As the complexity of educational goals increases and the technologies available to measure them evolve, the principles of PAD become not just beneficial but essential to creating assessments worthy of the time students and educators invest in them. The field may be approaching its inflection point; the question is whether we will be ready when it arrives.

References

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

CAST. (2018). *Universal Design for Learning Guidelines, version 2.2.*

Christensen, C. M. (1997). *The innovator's dilemma.* Harvard Business School Press.

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

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.

Grove, A. S. (1996). *Only the paranoid survive: How to exploit the crisis points that challenge every company.* Doubleday.

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

Hastings, R., & Meyer, E. (2020). *No rules rules: Netflix and the culture of reinvention.* Penguin.

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., Nichols, P., & Schneider, C. (in press). Designing and Developing Educational Assessments. In Linda L. Cook and Mary J. Pitoniak (Eds.), *Educational Measurement.* Oxford University Press.

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.

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/wpcontent/uploads/2019/02/SCILLSS_PLD_WhitePaper_V1812-02_FINAL_2_7_19.pdf

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

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

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

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

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

References

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.

Office of Elementary and Secondary Education. (2018, September 24). *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* U.S. Department of Education. https://www.ed.gov/sites/ed/files/2023/11/assessmentpeerreview.pdf

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., 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.

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.

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

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

Tsai, Y.-S., Whitelock-Wainwright, A., Chiu, Y.-L., He, Y., & Gašević, D. (2020). User experience design for technology-enhanced learning: A systematic review. *British Journal of Educational Technology, 51*(6), 2005–2033.

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

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

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

About the authors

Kristen Huff, M.Ed., Ed.D., currently serves as the Head of Measurement at Curriculum Associates, where she leads a team of assessment designers, psychometricians, and researchers in the development of online assessments integrated with personalized learning and teacher-led instruction. Prior to this role, she served as the Senior Fellow for the New York State Education Department as well as serving in leadership roles with several major assessment companies. Dr. Huff has deep expertise in k-12 large scale assessment, and has presented and published consistently in educational measurement conferences and publications for over 25 years. She served previously as a technical advisor for the 2026 NAEP Frameworks in Reading and Mathematics and as the inaugural Co-Chair of the NCME Task Force on Classroom Assessment 2016-2020. She was named as recipient of the 2021 Career Achievement Award from the Association of Test Publishers, and now serves as the NCME Representative to the Management Committee for the revision of the 2014 Joint Standards for Educational and Psychological Testing, published by AERA, APA, and NCME. Dr. Huff is first author of the forthcoming *Educational Measurement, 5th Edition* (Oxford University Press), and Designing and Developing Educational Assessments (Huff, Nichols, and Schneider).

About the Study Group

The Study Group exists to advance the best of artificial intelligence, assessment, and data practice, technology, and policy; uncover future design needs and opportunities for educational systems; and generate recommendations to better meet the needs of students, families, and educators.

Date of Publication
November 2025

Citation
Huff, Kristen (2025). Designing and Developing Educational Assessments for Contemporary Needs, *Handbook for assessment in the service of learning, Volume I: Foundations for assessment in the service of learning*. University of Massachusetts Amherst Libraries

Licensing
This case study is based on a chapter that has been made available under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND) license.