CASE STUDY

Arguments in Support of Innovating Assessments

James W. Pellegrino

Arguments in Support of Innovating Assessments

James W. Pellegrino

Abstract

This introductory chapter establishes assessment as a process of reasoning from evidence and presents the main arguments for why we need to innovate assessments, especially if they are to serve in support of learning. The first argument is that assessment should measure what matters, not just what is easy to measure. This means expanding the range of educational outcomes we assess to include the complex cognitive, socio-cognitive, and socio-emotional constructs that are essential for success in the worlds of today and tomorrow. The second argument is that we need new assessment design approaches and methods that leverage the affordances of digital technology to provide rich, meaningful, and useful sources of data and information. Following from the first two arguments, the third is that assessments should measure what matters and measure it well. Careful attention must be paid to the issues of validity and comparability when complex constructs are targeted for assessment, and when new tasks and tools are used for generating and interpreting evidence about student knowledge and skills.[1]

1 The structure and substance of this paper draws heavily from a previously published chapter by the author in the volume entitled Innovating Assessments to Measure and Support Complex Skills edited by Natalie Foster and Mario Piacentini and published by OECD in 2023.

Introduction

The Handbook for Assessment in the Service of Learning series continues arguments about assessment innovation and use that have been discussed in recent volumes such as *Innovating Assessments to Measure and Support Complex Skills* (Foster & Piacentini, 2023); *Classroom-based Assessment in STEM: Contemporary Issues and Perspectives* (Harris et al., 2023); and *Reimagining Balanced Assessment Systems* (Marion, Pellegrino, & Berman, 2024). The three volumes of this *Handbook* further contribute to those prior discussions by their collective attempt to tackle and broaden:

1) the "what" of assessment; 2) the "how" of assessment; and/or 3) the "value proposition" of assessment, i.e., the interpretation and use of results from innovative assessments but with a particular focus on for whom we measure and the interpretive value of the information obtained therein.

To develop and elaborate the three main arguments of the *What*, the *How*, and the *Value Proposition* we begin with a brief discussion of a fundamental conception about assessment, namely that it constitutes a process of reasoning from evidence guided by theory and research on critical aspects of the acquisition and development of knowledge and skill. We conclude this chapter with an additional argument of consequence for educational policy and practice—to achieve innovation in assessment and effect positive impact on educational outcomes, more coherent systems of assessment are needed. Such systems better connect assessments to one another given their intended interpretive uses regarding the constructs that matter and their relationship to curriculum and instruction, respectively.

Assessment as a Process of Reasoning from Evidence

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

The process of collecting evidence to support inferences about what students know and can do represents a chain of reasoning from evidence about student competence that characterises all assessments from classroom quizzes and standardised tests to computerised tutoring programmes, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. The first question in the assessment reasoning process is: "evidence about what?" *Data* become *evidence* in an analytic problem only when one has established their relevance to a conjecture being considered (Schum, 1987). Data do not provide their own meaning; their value as evidence can arise only through some interpretational framework. In the present context, educational assessments provide data such as written essays, marks on answer sheets, presentations of projects, or students' explanations of their problem solutions. These data become evidence only with respect to conjectures about how students acquire knowledge and skill.

In the *Knowing What Students Know* report (Pellegrino et al., 2001), the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the assessment triangle. The vertices of the *assessment triangle* represent the three key elements underlying any assessment (see Figure 1): a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence considering the assessment purpose and student understanding. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, or evaluated, without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the *Knowing What Students Know* report is that for an assessment to be effective and valid, the three elements must be in synchrony. The assessment triangle provides a useful framework for analysing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing their validity (e.g., see Marion and Pellegrino, 2006; Pellegrino et al., 2016).

The *cognition* corner of the triangle refers to theory, data, and a set of assumptions about how students represent knowledge and develop competence in an intellectual domain (e.g., fractions, Newton's laws, or thermodynamics) or regarding a socio-emotional skill or capacity. In any particular assessment application, a theory of competence in the domain of assessment is needed to identify the set of knowledge and skills that is important to measure for the intended context of use, whether that be to characterise the competencies students have acquired at some point in time to make a summative judgment, or to make formative judgments to guide subsequent instruction so as to maximise future learning. A central premise is that the theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain.

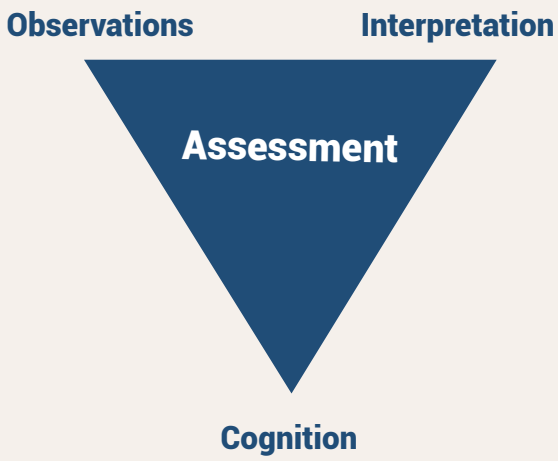
Every assessment is also based on a set of assumptions and principles about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates important knowledge and skills. The tasks to which students are asked to respond on an assessment are not arbitrary; they must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of

FIGURE 1: ***The assessment triangle***

Cognition
Theories, models & data about how students represent knowledge & develop competence in a domain of instruction and learning.

Observations
Tasks or situations that allow one to observe students' performance.

Interpretation
Methods for making sense of the evidence coming from students' performances.



Source: Pellegrino et al. (2001).

inferences and decisions that will be made based on the assessment results. The observation vertex of the assessment triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. In assessment, one has the opportunity to structure some small corner of the world to make observations. The assessment designer can use this capability to maximise the value of the data collected, as seen through the lens of the underlying assumptions about how students learn in the domain.

Every assessment is also based on certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterisation or summarisation of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher and is often based on an intuitive or qualitative model rather than a formal statistical one. Even informally, teachers make coordinated judgments about what aspects of students' understanding and learning are relevant, how a student has performed on one or more tasks, and what the performances mean about the state of a student's knowledge and understanding.

A crucial point is that each of the three elements of the assessment triangle not only must make sense on its own but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus, to have a valid and effective assessment, all three vertices of the triangle must work together in synchrony.

01

Argument 1: Measuring What Matters

Education research has well established that teachers, students and local and national policy makers take their cues about the goals for instruction and learning from the types of tasks found on state, national, and international assessments. Thus, what we choose to assess in areas such as science, mathematics, literacy, history, problem solving, collaboration, and critical thinking is what will end up being the focus of instruction. It is therefore critical that our assessments best represent the forms of knowledge and competency and the kinds of learning we want to emphasise in our classrooms if students are to achieve the complex, multi-dimensional proficiencies needed for the worlds of today and tomorrow. Doing so, however, requires that we move away from *measuring what is easy to measuring what matters*.

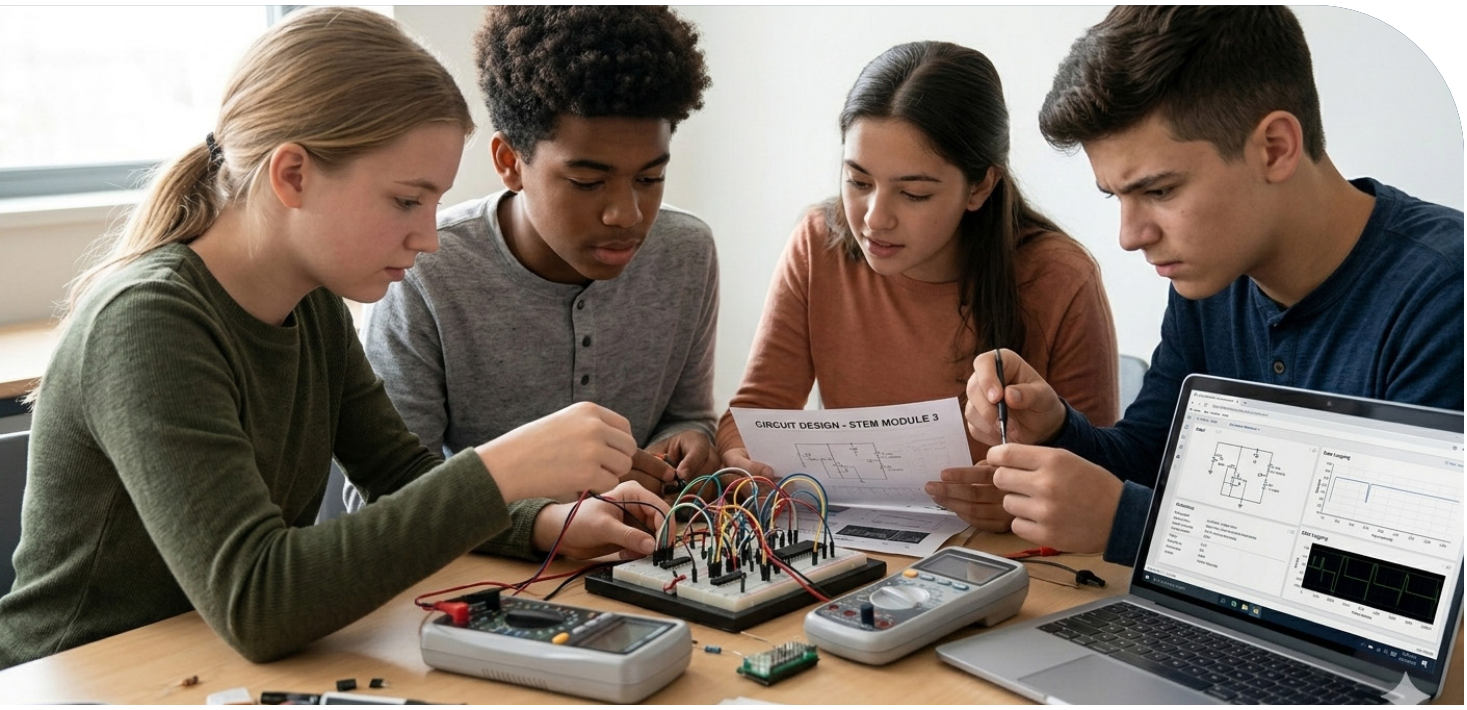
There is an increasing push to encourage students to develop "21st-century skills" that combine habits of mind and that include social and affective competencies (e.g., Bellanca, 2014; Pellegrino and Hilton, 2012). The European Commission's *Rethinking Education* (2012) reform effort emphasizes the need to promote transversal skills in education, such as critical thinking and problem solving. Additionally, PISA–the international assessment of student abilities administered by the OECD–has begun testing broader competencies that go beyond the disciplinary areas of mathematics, reading and science, such as problem solving and collaborative problem solving. Such 21st-century skills–or 21st-century competencies–are deemed necessary to prepare a global workforce to succeed in a new information-driven economy. Individuals must have the problem-solving, critical thinking, and collaboration and communication skills to evaluate and make sense of new information and to act upon this information in a range of settings.

Business leaders, educational organisations and researchers have begun to call for new education policies that target the development of such broad, transferable skills and knowledge. For example, the US-based Partnership for 21st-Century Skills (2010) argues that student success in college and careers requires four essential skills: critical thinking and problem solving, communication, collaboration, and creativity and innovation. The NRC Report *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st-Century* (Pellegrino

and Hilton, 2012) argued that the various sets of terms associated with the "21st-century skills" label reflect important dimensions of human competence that have been valuable for many centuries, rather than skills that are suddenly new, unique, and valuable today. The important difference across time may lie in society's desire for all students to attain levels of mastery–across multiple areas of skill and knowledge–that were previously unnecessary for individual success in education and the workplace. At the same time, the pervasive use of new digital technologies has increased the pace of communication and information exchange throughout society with the consequence that all individuals may need to be competent in processing multiple forms of information to engage in critical thinking and accomplish tasks that may be distributed across contexts that include home, school, the workplace and social networks (see e.g., Zlatkin-Troitschanskaia, Pellegrino, & Blatnik, in press).

To shift from policy into practice, assessments need to be able to measure these skills and competencies. To do that we need to have clear conceptions and definitions of the constructs to be assessed (the Cognition), the forms of evidence associated with those constructs (the Observations), and ways to make sense of that evidence for the purposes of reporting and use (the Interpretation).

Many of the *Handbook's* chapters explicitly focus on the "what" of educational assessment–the key constructs that we should be interested in assessing, why those constructs are important, and where we stand with respect to assessing them given the current educational assessment landscape. The bulk of the argument is that we should be focused on complex cognitive and socio-cognitive constructs, both within and across-disciplinary domains. The chapters discuss what we mean by these constructs and the types of tasks and situations where individuals would be required to exercise the requisite competencies, thereby providing the types of evidence that would be valid, interpretable and useful whether the intended use is at the classroom level to guide learning and instruction or in a large-scale educational monitoring context. Several chapters illuminate ways in which we might conceptualise and operationalise these constructs as well as some of the challenges in doing so. They set the stage for other chapters that move beyond conceptualisation of what we may want and need to assess as part of the advancement of 21st-century education, to the details of the design process and ways in which technology can enable the creation of situations that will provide the evidence we need while also assisting in the process of making sense of that evidence.



02

Argument 2: Assessment Design Processes and Applications of Technology

While it is especially useful to conceptualise assessment as a process of reasoning from evidence, the design of an actual assessment is a challenging endeavour that needs to be guided by theory and research about cognition as well as practical prescriptions regarding the processes that lead to a productive and potentially valid assessment for a particular context of use. As in any design activity, scientific knowledge provides direction and constrains the set of possibilities, but it does not prescribe the exact nature of the design, nor does it preclude ingenuity to achieve a final product. Design is always a complex process that applies theory and research to achieve near-optimal solutions under a series of multiple constraints, some of which are outside the realm of science. In the case of educational assessment, the design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a programme), the context in which it will be used (e.g., classroom or large-scale), and practical constraints (e.g., resources and time).

Recognising that assessment is an evidentiary reasoning process, it has proven useful to be more systematic in framing the process of assessment design as an Evidence-Centered Design process (e.g., Mislevy & Haertel, 2006; Mislevy and Riconscente, 2006). The process starts by defining the claims that one wants to be able to make about student knowledge and the ways in which students are supposed to know and understand some particular aspect of a content domain. Examples might include aspects of algebraic thinking, ratio and proportion, force and motion, heat and temperature, etc. The most critical aspect of defining the claims one wants to make for purposes of assessment is to be as precise as possible about the elements that matter and express these in the form of verbs of cognition that are much more precise and less vague than high-level cognitive, superordinate verbs such as know and understand. Example verbs might include compare, describe, analyse, compute, elaborate, explain, predict, justify, etc. Guiding this process of specifying the claims is theory and research on the nature of domain-specific knowing and learning.

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

The precision that comes from elaborating the claims and evidence statements associated with a domain of knowledge and skill pays off when one turns to the design of tasks or situations that can provide the requisite evidence. In essence, tasks are not designed or selected until it is clear what forms of evidence are needed to support the range of claims associated with a given assessment situation. The tasks need to provide all the necessary evidence and they should allow students to "show what they know" in ways that are as unambiguous as possible with respect to what the task performance implies about student knowledge and skill, i.e., the inferences about student cognition that are permissible and sustainable from a given set of assessment tasks or items.

In the *Knowing What Students Know* report (Pellegrino et al., 2001), many of the affordances of technology for advancing assessment design and practice were discussed in terms of the three interconnected components of the assessment triangle. The brief discussion that follows focuses on the constructs that could be represented in innovative assessment frameworks (Cognition), the ways in which those constructs could be realised in the assessment environment (Observation), and some of the interpretive challenges and solutions associated with doing so for purposes of measurement and reporting (Interpretation).

The Cognition vertex of the assessment triangle

What matters in assessment is what we are trying to reason about–the contemporary conception of student Cognition in a domain that matters to domain experts, educators and society. As the conception of student cognition changes and expands in terms of what students are supposed to know and be able to do, as has been the case for many domains, technology affords opportunities for substantially changing and extending the *Observation and Interpretation* components of the assessment triangle to more adequately represent and provide evidence about the constructs of interest. Doing so enhances the entire evidentiary reasoning process and the validity of an assessment given its intended interpretive use.

The Observation vertex of the assessment triangle

Technology provides opportunities for the presentation of dynamic stimuli (e.g., videos, graphics, 2- and 3-D simulations) that can be interacted with in the service of eliciting relevant sets of responses from students. Simultaneously, technology enables the generation and capture of a variety of response products, including situations in which students generate responses using multiple modalities (e.g., drawing and writing). Technology-enhanced assessments enable engagement with a variety of content and practices by opening the door to interactive stimulus environments and response formats that better match the intended reasoning and response processes that form the basis for desired claims about student proficiency (Gorin and Mislevy, 2013).

Students' interactions with these technology-enhanced assessments can be logged to provide data on how they engage in particular processes. For various 21st-century competencies, the process by which one completes the activity can be as important a piece of information about knowledge and skill as the final product. In these cases, understanding the operations that students performed in the process of creating the final product may be critical to evaluating students' proficiency. Log data offer the opportunity to reveal these actions, including where and how students spend their time, and what choices they make in situations like using a simulation. Such applications offer the potential to provide large volumes of "clickstream" and other forms of response process data that might be useful for making inferences about student thinking (Ercikan and Pellegrino, 2017).

The Interpretation vertex of the assessment triangle

Technology offers significant opportunities to enhance the reasoning-from-evidence process given the types of observations described above. Collecting these types of data makes little sense unless there are ways to reliably and meaningfully interpret them. This can evolve through mechanisms such as automated scoring of responses and application of complex parsing, statistical and inferential models for response process data (see Ercikan and Pellegrino, 2017). Critical data to consider include the time taken to perform various actions, the actual activities chosen, and their sequence and organisation. The potential exists for examining the global and local strategies students use while solving assessment problems and their implications, including how such strategies relate to the accuracy or appropriateness of final responses. Although capturing such data in a digital environment is "easy," making sense of the data is far more complicated.

The same can be said for capturing data to constructed response questions where students may be expressing in written and/or graphical form an argument or explanation about some social, economic or scientific problem or phenomenon, describing the design of an investigation, or representing a model of some structure or process.

The data capture contexts described above are challenging regarding scoring and interpretation. It is here that artificial intelligence and machine learning may play a significant role in future innovative assessments (see e.g., Zhai et al., 2020a,b). Developments in machine learning also may allow researchers to analyze complex response process data, including to reveal patterns that provide important insights into students' cognitive processes in problem solving (Zhai et al., 2020a, 2020b, 2021a, 2021b; Zhai, 2021). Such data may prove to be especially informative about student thinking and reasoning and thus add greatly to the knowledge gained about student competence from large-scale assessments like PISA. An interesting example was provided in a recent report by Pohl et al. (2021) who showed that differences in student response processes, when combined with scoring methods, can significantly change the interpretation of a country's performance in PISA.

In summary, digital technologies hold great promise for helping to bring about changes in assessment that many believe are necessary. Technologies available today and innovations on the immediate horizon can be used to access information, create simulations and scenarios, allow students to engage in learning games and other activities, and enable collaboration among students. Such activities make it possible to observe, document and assess students' work as they are engaged in natural activities–perhaps reducing the need to separate formal, external assessments from learning in the moment (e.g., Behrens, DiCerbo, and Foltz, 2019). Technologies will certainly make possible the greater use of formative assessment that in turn has been shown to significantly impact student achievement. Digital activities may also provide information about abilities such as persistence, creativity and teamwork that current testing approaches cannot. Juxtaposed with this promise is the need for considerable work to be done on issues of scoring and interpretation of evidence before such embedded assessment can be useful for these varied purposes. Suffice it to say that the technology and assessment field is advancing at a very rapid rate and providing potential solutions to many of the concerns and possibilities noted above. Advances in assessment are increasingly being influenced by the rapid advances in artificial intelligence and data analytics (see e.g., multiple chapters in the volumes edited by Foster & Piacentini, 2023 and by Zhai & Krajcik, 2024; as well as Zhai & Wiebe, 2023).

Developing assessments of complex cognitive competencies requires being explicit about all three elements of the assessment triangle and their inter-relationships. Multiple chapters in the Handbook address various aspects of Argument 2 regarding the observation and interpretation elements of the assessment triangle, with an emphasis on how technology can be exploited through and within a principled design process to create assessments of the complex cognitive and socio-cognitive performances that matter. Through a combination of argument and specific examples, these chapters provide support for the claim that next-generation assessments are possible but can only be generated through a highly principled design process that makes explicit the evidentiary chain of reasoning at the core of valid assessment. The chapters also reveal the complexities that accrue in designing such assessments and then making sense of the multiple forms of evidence they can produce.

Argument 3: Valid Interpretation and Use of Results

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

Kane (2006) and others (Haertel and Lorie, 2004; Mislevy et al., 2003) distinguish between: 1) the interpretive argument, i.e., the propositions that underpin test score interpretation; and 2) the evidence and arguments that provide the necessary warrants for the propositions or claims of the interpretive argument. In essence this view identifies as the two essential components of a validity argument the claims being made about the focus of an assessment and how the results can be used (interpretive argument), together with the evidence and arguments in support of those claims. Appropriating this approach, contemporary educational measurement theorists have framed test validity as a reasoned argument backed by evidence (e.g., Kane, 2006). An argument and evidence framing of validity supports investigations for a broad scope of assessment designs and purposes, including many that go beyond typical large-scale tests of academic achievement or aptitude and move one into the arena of innovative and instructionally supportive assessments (e.g., see Pellegrino et al., 2016).

Given the nature of the constructs of interest, including their inherent complexity and multi-dimensionality, we must acknowledge from the outset the challenges that will be faced in establishing validity arguments for innovative assessments of 21st-century competencies, including the reporting of results for various intended use cases. Validity arguments will depend on well-developed interpretive arguments that include: 1) clear specifications of the constructs of interest and their associated conceptual backing; 2) the forms of evidence associated with those constructs; and 3) the methods for interpretation and reporting of that evidence. Such interpretive arguments are essential to guide assessment design processes, including carefully thought-out applications of technology and data analytics to support the observational and inferential aspects of the overall reasoning from the evidence process. As noted above, carefully developed and articulated claims about what is being assessed and reported then need to be supported by empirical evidence. Such evidence can be derived from multiple forms of data involving variations in human performance and are essential to establishing an assessment's validity argument.

Assessing 21st-century competencies demands more than innovation—it requires rigorous validity arguments, strong empirical evidence, and an unwavering commitment to equity and fairness across diverse learners and contexts.

In pursuing innovative assessment of 21st century competencies, of paramount concern are issues of equity and fairness as part of the validity argument. Of particular concern is comparability of results and validity of inferences derived from performance obtained across different modes of assessment, especially for varying groups of students (see Berman et al., 2020). As assessment has moved from paper-and-pencil formats to digitally-based assessment, the general focus has been on mode comparability and concerns about student familiarity and differential access to the hardware and software used (see Way and Strain-Seymour, 2021). However, as the digital assessment world advances, a significant issue for innovative assessment is determining how student background characteristics including language, culture, and educational experience influence performance on different types of tasks and innovative assessment designs that leverage the power of technology. As the assessment environments and tasks become more innovative, equity and fairness concerns become even more important than general mode comparability effects. Thus, a key part of the validity argument for any innovative assessment will be establishing the socio-cultural boundaries related to equitable and fair interpretations and uses of the assessment results.

Many of the *Handbook's* chapters focus on critical aspects of design and development as part of establishing the validity of next-generation assessments for 21st-century competencies. More specifically, multiple chapters focus on the validity evidence that would be derived through the application of a principled design process that forces one to articulate, in varying degrees of detail, the connections between and among the cognition, observation and interpretation components of the assessment. Such evidence contributes to the assessment's overall validity argument but needs to be complemented by various forms of empirical data on how the assessment performs.

Towards More Coherent and Instructionally Supportive Systems of Assessment

No single assessment can evaluate all the forms of knowledge and skill that we value for students; nor can a single instrument meet all the goals held by parents, practitioners and policymakers. As argued below, it is important to envision a coordinated system of assessments in which different tools are used for different purposes–for example, formative and summative, or diagnostic vs. large-scale reporting. Within such systems, however, all assessments should faithfully represent the constructs of interest, and all should model good teaching and learning practice.

At least four major features define the elements of assessment systems that can fully reflect rigorous standards and support the evaluation of deeper learning (see Darling-Hammond et al. (2013) for an elaboration of the relevance, meaning and salient features of each of these criteria):

- *Assessment of higher-order cognitive skills* through most of the tasks that students encounter–in other words, tasks that tap the skills that support transferable learning, rather than emphasising only those that tap rote learning and the use of basic procedures. While there is a necessary place for basic skills and procedural knowledge, it must be balanced with attention to critical thinking and applications of knowledge to new contexts.
- *High-fidelity assessment of critical abilities*, as articulated in the standards–such as communication (speaking, reading, writing and listening in multi-media forms), collaboration, modelling, complex problem solving and research, in addition to key subject matter concepts. Tasks should measure these abilities directly as they will be used in the real world rather than through a remote proxy.

- *Use of items that are instructionally sensitive and educationally valuable*—in other words, tasks should be designed so that the underlying concepts can be taught and learned, distinguishing between students who have been well- or badly-taught rather than reflecting students' differential access to outside-of-school experiences (frequently associated with their socio-economic status or cultural context) or interpretations that mostly reflect test-taking skills. Preparing for (and sometimes engaging in) the assessments should engage students in instructionally valuable activities, and results from the tests should provide instructionally useful information.
- *Assessments that are valid, reliable, and fair* for a range of learners, such that they measure well what they purport to measure, be accurate in evaluating students' abilities and do so reliably across testing contexts and scorers. They should also be unbiased and accessible and used in ways that support positive outcomes for students and instructional quality.

A major challenge is determining the conditions and resources needed to create coherent systems of assessments that work across contexts ranging from the classroom to larger organisational units such as districts, states, countries and internationally. Regardless of their context of implementation, assessments in such systems must support the ambitious goals we have for the educational system, meet the information needs of different stakeholders, and align with the criteria above. The volume *Reimagining Balanced Assessment Systems* (Marion, Pellegrino, & Berman, 2024) provides a very powerful and comprehensive argument for such coherence with explicit principles for design and implementation across multiple levels of the educational system. In such balanced assessment systems all assessments are based on contemporary theory and research on knowing, learning and human development and all are focused on providing information that supports equitable and ambitious classroom teaching and learning.

Final Thoughts

Innovation and change are always challenging no matter the context. They have been especially challenging in education systems given long-standing and entrenched histories of educational policy and practice. Many have argued that education has changed little over the last 50–100 years in terms of how it is organised, delivered, what is taught and how it is assessed. Yes, there have been changes in the subject matter learned, in the pedagogies employed and, most recently, in the uses of technology. Those changes have been evolutionary and not revolutionary. Not surprisingly, much the same can be argued about educational assessment regarding what we assess and how we do so, including applications of technology to the practice of assessment.

This *Handbook* is focused on an alternative and perhaps revolutionary vision that starts with the complex competencies that are deemed critical for citizens of the 21st-century. The Handbook's chapters provide a vision of what they are by characterising how we might create environments and situations where the competencies of interest would necessarily be expressed in addition to describing the evidence that those environments could provide about those competencies. Some might find it curious that a vision for the future of education starts with assessment rather than curriculum and instruction. One of the benefits of thinking first about the outcomes we desire from the educational system, with a particular focus on what they would look like, is that this information provides the basis for a "Backwards Design" process regarding the design of curriculum and instruction that can lead to those outcomes (Wiggins and McTighe, 2011).

We hope they help you consider the costs and benefits of innovative educational assessment. These considerations include the competencies described, the types of environments for assessing them, conceptual and operational design and implementation challenges, and the value of the information derived in terms of its utility for classroom teaching and learning and for education more broadly. We also suggest that you consider what it might take to move in the directions highlighted by this volume given the many entrenched assumptions, policies and practices that have come to dominate the educational assessment landscape.

References

(AERA/APA/NCME) American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999, 2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Behrens, J. T., DiCerbo, K. E., & Foltz, P. W. (2019). Assessment of complex performances in digital environments. *The Annals of the American Academy of Political and Social Science, 683*(1), 217–232.

Bellanca, J. (2014). *Deeper learning: Beyond 21st-century skills.* Bloomington, IN: Solution Tree Press.

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (Eds.). (2020). *Comparability of Large-Scale Educational Assessments: Issues and Recommendations.* National Academy of Education.

Darling-Hammond, L., Herman, J., Pellegrino, J. W., et al. (2013). *Criteria for high-quality assessment. Stanford, CA: Stanford Center for Opportunity Policy in Education.*

Ercikan, K., & Pellegrino, J. W. (2017). *Validation of score meaning for the next generation of assessments: The use of response processes.* Taylor & Francis.

European Commission. (2012). *Rethinking education: Investing in skills for better socio-economic outcomes.*

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills.* Paris: OECD Publishing.

Gorin, J. S., & Mislevy, R. J. (2013). Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment [Paper presentation].

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

Harris, C., Wiebe, E., Grover, S., & Pellegrino, J. W. (Eds.). (2023). *Classroom-based assessment in STEM: Contemporary issues and perspectives.* Community for Advancing Discovery Research in Education (CADRE). Education Development Center, Inc.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

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

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

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

Marion, S., Pellegrino, J. W., & Berman, A. (Eds.). (2024). *Reimagining balanced assessment systems.* National Academy of Education.

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

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20.

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

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

Partnership for 21st-Century Skills. (2010). *21st-century readiness for every student: A policymaker's guide.* Tucson, AZ: Author. Available: https://files.eric.ed.gov/fulltext/ED519425.pdf

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

References

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. https://doi.org/10.1080/00461520.2016.1145550

Pellegrino, J. W., & Hilton, M. (Eds.). (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st-century.* National Academies Press.

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science, 372*(6540), 338–340.

Schum, D. (1987). *Evidence and inference for the intelligence analyst.* University of America Press.

Way, D., & Strain-Seymour, E. (2021). *A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress.* NAEP Validity Studies Panel.

Wiggins, G., & McTighe, J. (2011). *The understanding by design guide to creating high-quality units.* ASCD.

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology, 30*(2), 1–11.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020a). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459.

Zhai, X., Haudek, K. C., Wilson, C., & Stuhlsatz, M. (2021a). A framework of construct-irrelevant variance for contextualized constructed response assessment. *Frontiers in Education, 6,* 751283. https://doi.org/10.3389/feduc.2021.751283

Zhai, X., & Krajcik, J. (Eds.). (2024). *Uses of AI in STEM education.* Oxford University Press.

Zhai, X., Krajcik, J., & Pellegrino, J. (2021b). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology, 30*(2), 298–312.

Zhai, X., & Wiebe, E. (2023). Technology-based innovative assessment. In C. Harris, E. Wiebe, S. Grover, and J. W. Pellegrino (Eds.), *Classroom-based STEM assessment: Contemporary issues and perspectives,* (pp. 99–126). Community for Advancing Discovery Research in Education (CADRE). Boston: Education Development Center.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020b). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151.

Zlatkin-Troitschanskaia, O., Pellegrino, J. W., & Bartnik, T. (in press). Learning to think critically. In R. E. Mayer, P. A. Alexander, & L. Fiorella (Eds.), *Handbook of research on learning and instruction.* New York: Routledge.

About the author

James W. Pellegrino is Emeritus Professor of Psychology and Learning Sciences and Founding co-director of the Learning Sciences Research Institute at the University of Illinois Chicago. His research and development interests focus on children and adults thinking and learning and the implications of cognitive research and theory for assessment and instructional practice. He has published over 350 books, chapters, and articles on cognition, instruction, and assessment. His education research has been funded by the National Science Foundation, the Institute of Education Sciences, and private foundations. As Chair or Co-Chair of several National Academy of Sciences study committees he co-edited major synthesis reports on teaching, learning, and assessment, including Knowing What Students Know: The Science and Design of Educational Assessment. He previously served on the Board on Testing and Assessment of the National Research Council and is a lifetime member of both the National Academy of Education and the American Academy of Arts and Sciences. His service includes the Technical Advisory Committees of several states and consortia, as well as those of the College Board, ETS, OECD, and the National Center on Education and the Economy. He currently serves on the NCEE Board of Trustees and ETS' Research Advisory Council.

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
February 2026

Citation
Pellegrino, J. W. (2025). Arguments in support of innovating assessments. In E. M. Tucker, E. Armour-Thomas, & E. W. Gordon (Eds.), *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.