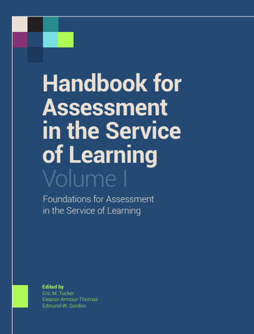
# Innovating Assessment Design to Better Measure and Support Learning

Natalie Foster and Mario Piacentini

UMassAmherst
University Libraries




**Series Editors:**
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Foster, N., & Piacentini, M. (2025). Innovating assessment design to better measure and support learning. 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.

# Innovating Assessment Design to Better Measure and Support Learning

Natalie Foster and Mario Piacentini



**Abstract**
Assessments should support deep learning as well as measure its outcomes. This chapter proposes five design principles–following insights from research in the learning and cognitive sciences–for designing innovative assessments to measure and support the development of complex competencies. These principles include: 1) using extended performance tasks; 2) accounting for students' prior knowledge; 3) providing opportunities for productive failure; 4) providing feedback and instructional support; and 5) designing "low floor, high ceiling" tasks. Our fundamental argument driving these design principles is that assessments should not only measure student progress towards educational goals, but also model and provide insight into students' deep learning processes. Adopting these design principles will enable educators to collect information about how well students can engage in complex thinking and problem-solving processes while reducing the current distance between assessment and learning. We illustrate how these design principles can be applied to assessments in different contexts and for different purposes, including a large-scale international assessment (the PISA 2025 Learning in the Digital World assessment) and a classroom-based formative assessment. We conclude by discussing areas for future research and reflecting on the different forms of capital (political, intellectual and financial) required for advancing this vision of assessment.

In recent years, researchers have argued for the need to innovate educational assessments to better measure and support the development of important skills[1] (e.g., Foster & Piacentini, 2023; Kyllonen & Sevak, 2024; Schwartz & Arena, 2013). Such works respond to shifts in educational discourse and policy about what is important to teach and learn–so-called 21st-century competencies. Innovation is required across the entire assessment development process: from conceptual foundations (defining the components of what are often complex constructs and the authentic contexts in which they are engaged), to design considerations (how tasks are designed and delivered to test takers), measurement issues (how to generate, interpret and accumulate useful evidence about what students know and can do), and reporting options (how to clearly communicate the results to intended audiences, be they teachers, learners, administrators or policy makers). To achieve positive impact on educational outcomes, more coherent systems of assessment are also needed, increasing the alignment between formative and summative assessment (Darling-Hammond et al., 2013; Pellegrino, 2023).

In this chapter, we set forth some key arguments in favor of innovating assessments to better measure and support students' learning. We then focus in on the design considerations that should guide the development of next-generation assessments that intend to measure and support the cultivation of complex skills. We propose five design principles that respond to key insights from research in the learning and cognitive sciences. These are: 1) use extended performance tasks; 2) account for student knowledge in task design and performance interpretation; 3) provide opportunities for productive failure; 4) provide feedback and instructional support during tasks; and 5) design "low floor, high ceiling" tasks. One of the key arguments of our approach is that assessments should not only be useful for measuring student progress towards educational goals, but they should also model and provide insight into students' deep learning processes (Piacentini et al., 2023). In this way we can collect important data on complex thinking and problem-solving processes, and at the same time, we can reduce the current distance between assessment and the learning we want to promote in classrooms. We illustrate how these design principles can be applied to assessments in different contexts

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

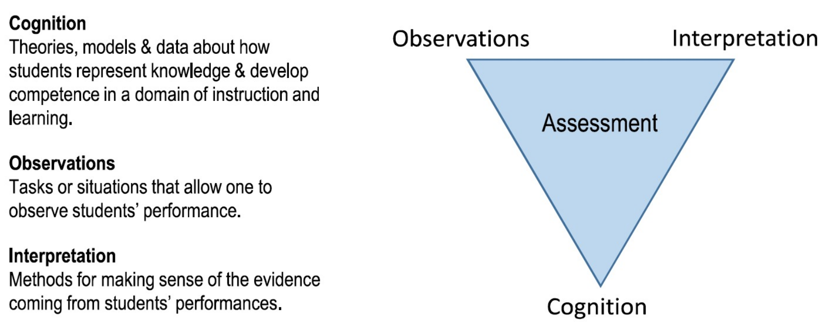
and for different purposes, including a large-scale international assessment (the Learning in the Digital World assessment, to be conducted in the 2025 cycle of the Programme for International Student Assessment) and a classroom-based formative assessment.

### Arguments in support of innovating assessments

A fundamental conception about assessment is that it constitutes a process of reasoning from evidence, guided by theory about the critical aspects of knowledge and skill one is interested in measuring. This process of reasoning from evidence has been portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino et al., 2001). The vertices of the assessment triangle represent the three key elements underlying any coherent assessment: a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence (See Figure 1).

**Figure 1**

***The assessment triangle, Source: Pellegrino et al. (2001).***




The assessment triangle provides a useful framework for analyzing the foundations of current assessments to determine how well they accomplish the educational goals we value, as well as how we can design future assessments and establish their validity (Marion & Pellegrino, 2007; Pellegrino et al., 2016). Before we discuss how future assessments might be designed to better support learning and instruction, we first consider the extent to which current assessments are fit for purpose given an increasing focus on preparing students for the future equipped with so-called 21st-century competencies. We also briefly discuss some of the challenges associated with developing assessments of more complex skills and competencies as well as new opportunities for innovation in test design and measurement technology.

***Measuring what matters***

It is no surprise that what we choose to assess ends up being the focus of instruction. By the same logic, those skills and constructs that we do not assess inevitably fall by the wayside in terms of instruction focus. Since the turn of the century, business leaders, educational organizations and researchers have increasingly called for new education policies targeting the development of broad, transferable skills and knowledge in education to prepare today's students for participation in an emerging global knowledge society. Information and communication technologies (ICTs) have radically transformed our societies, connecting people around the world, delivering unprecedented amounts of information to us, and giving rise to new forms of decentralized and autonomous learning. Young people today must not only learn to participate in a more interconnected, digital and rapidly changing world, they must also develop agency.

Several international frameworks have focused on concretely identifying the sets of knowledge, skills and attitudes–or so-called 21st-century competencies–that young people require to succeed in their futures (Binkley et al., 2011; European Commission, 2019; Fadel & Groff, 2018; OECD, 2018; Pellegrino & Hilton, 2012; Scott, 2015; World Economic Forum, 2015). Broadly speaking, some distinct categories of competencies emerge across these different frameworks including cognitive (e.g., creative thinking, critical thinking, problem solving), interpersonal (e.g., communication, collaboration), metacognitive (e.g., self-regulated learning), intrapersonal (e.g., persistence, adaptability) and digital (e.g., media literacy, digital literacy) (Foster, 2023). Digital competencies are often a central component that intersect with other 21st-century competencies given the vast proliferation of ICTs and related advances (e.g., Artificial Intelligence) in both our personal and professional lives, in turn transforming our individual and collective capacities for communication, collaboration, searching for information and using knowledge.

Yet for as much as 21st-century competencies are now a familiar component of contemporary discourse on education, assessments need to be able to measure these competencies if we want to truly shift policy into practice. It is critical that our assessments reflect the forms of knowledge and learning we want to emphasise in our classrooms if students are to achieve the complex and multi-dimensional competencies they will need for the worlds of today and tomorrow. Not only does this mean being able to measure the extent to which students master these important competencies in different contexts, but also designing assessments that can provide timely feedback to learners and educators that can benefit instructional practices. In other words, developing assessments that *support* deep learning as well as measure its outcomes. However, as argued by Pellegrino (2023), doing so requires that we move away from *measuring what is easy* to *measuring what matters.*

Of course, measuring what matters is easier said than achieved. Several interconnected conceptual and practical challenges exist when designing assessments of 21st-century competencies, particularly in the context of large-scale assessment. These challenges include defining the assessment construct and learning progressions for complex competencies at different levels of proficiency, identifying the extent to which domain-specific knowledge supports performance and performance in an assessment can be generalized to other contexts, being able to identify and design tasks that elicit both outcomes and the processes that often define these competencies, and being able to generate interpretable evidence about students' ways of thinking and doing (see Foster, 2023, for an in-depth discussion of each). While these challenges remain real and require significant investments of time, intellectual and financial resources to overcome, some promising examples of assessments of 21st-century competencies exist at scale (see the OECD's Programme for International Student Assessment (PISA) or the Assessment and Teaching of 21st-Century Skills (ACT21S) project). What's more, advances in computer-based assessment–and particularly design and measurement technology–provide new opportunities to pursue this assessment agenda.

### *Unlocking new advances in design and measurement technology*

Advances in technology offer much potential to transform what we can measure, how we can make sense of test taker performance, and how assessment can relate to learning (Thornton, 2012; Shute & Kim, 2013; Timmis et al., 2016; Zhai et al., 2020; DiCerbo et al., 2017). Technology can introduce new forms of active, immersive and iterative performance-based tasks in interactive environments that make it possible to observe how test takers engage in authentic, open-ended learning and problem-solving activities. These can provide richer observations and potential evidence about students' thinking processes, behaviors and decision-making, as well as enable the measurement of dynamic skills beyond the capability of more traditional and static test items. These environments allow students to engage actively in the processes of making and doing, making it possible to track the strategies that test takers employ and the decisions they make as they work through complex tasks that evolve on their basis of their actions. Such student-centred tasks also provide opportunities for students to learn and develop skills as they make decisions and engage in iterative problem-solving. In addition to providing dynamic problems to test takers, such technology-enhanced environments can also replicate certain situations in a standardized way for assessment (e.g., simulating collaborative encounters) or visualize scenarios that would otherwise be impossible in school settings (e.g., modelling dynamic systems).

New task modalities, problem types and affordances mean there are also new possibilities for response types–and in turn, new sources of potential evidence about test takers (Sabatini et al., 2023). Digital platforms can capture, time stamp and log student interactions with the test environment. This is especially transformative in the context of measuring 21st-century competencies, as the process by which an individual engages with an activity can be just as valuable for evaluating proficiency as their final product. These process data, when coupled with appropriate analytical models, can reveal how students engage with problems, the choices they make, and the strategies they do (or do not) implement. Patterns of behavior associated with different mastery levels can be identified, which can then be used to augment the precision of performance scores and provide diagnostic information to educators about students.

Taken together, these opportunities represent a powerful shift in assessment away from simple knowledge reproduction to knowledge-in-use–exactly the types of performances in which 21st-century competencies are engaged and developed. Measuring these constructs well necessarily requires assessments to be closer tied to the processes and contexts of learning and instruction–something that technology can facilitate through embedded assessment. Embedded assessment environments can integrate pedagogical affordances like scaffolding and feedback to explicitly support learning, merging summative and formative assessment and providing measures of the capacity of test takers to learn and transfer their learning to other tasks. By allowing for the real-time measurement of students' capacities as they engage in meaningful learning activities, technology holds the promise of creating new systems of evaluation where evidence on students' progress is collected in a continuous and unobtrusive way. In turn, educators can gain better insights into their students' learning processes, and assessment and learning are no longer explicitly separated activities.

**Design principles for innovative assessments**

While these opportunities are exciting, few extant assessment systems take full advantage of this potential to measure what matters. Traditional assessments (especially large-scale) have evolved to comply with practical constraints on testing time, cost efficiency and established measurement models; this often resulting in a narrow focus on knowledge mastery and response correctness, abstracting from the cognitive, interpersonal and intrapersonal processes that support test taker responses. We argue that test taker performance in the types of restricted and static situations used in traditional assessments might not provide particularly valid inferences on their capabilities to think, act and learn in real-life situations.

While the technology-enabled innovations described above clearly offer new possibilities for assessment design that can bridge this gap, it's important to acknowledge that such innovations are only useful insofar as they are integrated purposefully within a principled design process. This means that choices about what aspects of performance to simulate, what tools and affordances to include, what evidence to collect, and what interpretations to draw from the data are guided by an explicit chain of reasoning. In the rest of this chapter, we contend that it is possible to address this misalignment between current practice and promise in two ways: first, by taking stock of research that has investigated the mental structures that support the types of learning and problem solving encompassed by 21st-century competencies; and second–much more difficult–by creating an internally consistent system of teaching practices and assessments that reflect these research insights. In this system, deeper learning experiences prepare students for future learning, and innovative assessments measure how effectively students have engaged with these deeper learning experiences.

### *Insights about deeper learning processes and their implications for assessment*

One of the main conclusions from research in the learning sciences is that learning is a socially situated process (Dumont et al., 2010; Darling-Hammond et al., 2019). People learn and develop expertise by participating in the practices of a community of experts and by learning to use the tools, languages and strategies that have been developed within that community (Mislevy, 2018; Ericsson, 2006; Pellegrino & Hilton, 2012). Becoming skillful in any domain involves learning the ways of thinking and acting that are aligned with those valued by a community of experts and by soliciting and using their feedback. That learning is mediated by socially constructed practices has clear implications for instruction: deeper learning occurs when students engage in activities that are realistic, complex, meaningful, and motivating, and when they can call upon the experience of knowledgeable others for guidance and support. Assessment experiences aimed at engaging and measuring deeper learning processes must therefore replicate these features.

Some research has focused on contrasting "experts" (i.e., individuals that have constructed mental models in a given domain and who, through participation in key practices, have learned to apply these ideas to solve new problems) with "novices" (i.e., those who have not consolidated their basic knowledge in a domain through practice). A recurrent observation is that experts have strong metacognitive skills (Hatano, 1990). While learning and problem solving, they engage in regulatory behaviors such as knowing when to apply a procedure, planning, predicting the outcomes of an action, questioning the limitations of their knowledge, monitoring their progress, and efficiently apportioning cognitive and emotional resources. The capacity to regulate one own's learning and adapt accordingly further distinguishes routine experts from adaptive experts, with the latter being "characterized by their flexible, innovative, and creative competencies within the domain" (Hatano & Oura, 2003, p. 28). Instructional and assessment practices that aim to develop and differentiate experts and novices must provide learners with opportunities to engage these regulatory behaviours.

Research has also shown that general problem-solving procedures ("weak methods"), such as trial-and-error or hill climbing, are slow and inefficient (Pellegrino et al., 2001), and that experts instead use deep knowledge of the domain ("strong methods") to solve problems. This deeper knowledge does not refer to isolated facts, but rather to knowledge encoded in a way that closely links it with its contexts of practice and conditions of use. When experts face new problems or scenarios, they can readily activate and retrieve the subset of their knowledge that is relevant to the task at hand (Simon, 1980). Learners progress in their mastery of a discipline through similar processes of acquisition and use of increasingly well-structured knowledge schema in different contexts. The implication for assessment here is that learning and skill mastery cannot readily be divorced from their specific contexts of practice, and as such, measuring deeper learning processes must occur in context rather than in highly generalised problem contexts.

### *Design innovations to measure learning and support teaching*

Thus far, we have described the types of competencies that students need to learn, problem-solve, and ultimately thrive in their future roles within society. Thanks to extensive research in the learning sciences, we also increasingly understand aspects of the processes involved in developing these higher order thinking and learning skills, and the types of experiences that elicit them. In this section, we highlight five broad design principles that we argue should form the basis of “next-generation assessments” that can yield potentially valid evidence about where students are in the development of 21st-century competencies (in summative applications) and about what they need to do to progress in these skills (in formative applications). For a more detailed discussion and examples of these principles, see Piacentini et al. (2023).

These five design principles also closely connect to the Principles set forth in and throughout this Handbook, namely that the focus of assessment should target and consider progress, outcomes and processes that can be transferred to other settings, situations and conditions (Principle 2), and that the design of assessments should support learners’ processes, motivation, attention, engagement, effort and metacognition (self-regulation) (Principle 4).

### *Design Principle 1: Include Extended Performance Tasks*

Assessments that aim to measure how prepared students are for deeper learning–and to generate insights useful to teaching and learning processes–must engage students in active and authentic learning processes. Students engage 21st-century competencies in situations in which they interact with others, evaluate available resources, make choices about the course of action to take, try out strategies, iterate, and adapt according to the results. From an assessment perspective, this means providing students with a purposeful challenge that replicates the key features of those educational experiences where deeper learning happens. Evaluating students' capacity to construct new knowledge in choice-rich environments means that students should be given sufficient time and affordances to demonstrate what they can do. We argue that extended units with multiple activities, sequenced as steps towards achieving a main learning goal, can provide a more authentic and motivating experience of assessment that provides valid data about students' competencies (i.e., evidence that is predictive of what students can do outside of the constrained and stressful context of a test).

Advances in technology now allow much more data to be captured on how students spend their time in extended tasks by immersing them in simulated environments and communities of practice. These environments can facilitate a more open interpretation of students' goals and their exploration of the problem constraints, reward diverse solution strategies and outcomes, and provide feedback to learners. This also makes it possible to observe metacognitive processes that are crucial to learning in a non-obtrusive way, tracking how students plan and implement strategies, how they behave when they are stuck, and how they respond to feedback (Nunes et al., 2003). The application of a principled design process can lead to a productive use of these process data to augment the evidence that is derived from final solutions, therefore reducing the trade-off between reliability and authenticity (e.g., Piacentini, 2023; Sabatini et al., 2023).

### *Design Principle 2: Account for Knowledge in Task Design and Reporting of Performance*

Knowledge plays an important role in authentic task performance. Competencies like creativity, critical thinking or communication are rarely exercised within a vacuum. In an assessment context, students' ability to demonstrate these skills will always be observed within a given context and their knowledge about this context or situation will influence the type of strategies they use as well as what they are able to accomplish (Mislevy, 2018). Attempting to design completely decontextualized assessment problems or scenarios threatens validity: if no knowledge is required to solve a task, can an assessment really measure the types of complex competencies it claims to be interested in?

What this means, in turn, is that it is important to explicitly identify the knowledge that students need in any assessment context to meaningfully engage with the test activities and to evaluate the extent to which differences in prior knowledge influence the evidence we can obtain on the target constructs. It is also important to assess complex skills across a variety of knowledge and application domains to make valid conclusions on students' mastery of these skills. Evaluating the extent to which students possess relevant knowledge when engaging with a complex performance task (for example, through a short battery of items) should become an integral part of the design and assessment process in next-generation assessments, as this can help to interpret their subsequent behaviors and choices during the assessment (Piacentini, 2023; Roll & Barhak-Mirkowitz, 2023).

### *Design Principle 3: Provide Opportunities for Productive Failure*

There is evidence that we can make robust claims about students' preparedness to learn new things by studying how they work on unfamiliar problems (Roll et al., 2011; Schwartz & Martin, 2004). For example, "invention activities" ask students to work on problems requiring concepts or procedures that they have not yet been taught, with the aim that students explore and understand the core properties of a construct before being taught expert solutions and strategies (Roll et al., 2012). Students often fail in their attempts to solve or generate canonical solutions to such problems, but experimental evidence shows that students who learn through invention activities are better at transferring their knowledge (i.e., solving other tasks requiring the same knowledge schemes in a different context) in comparison to students who are directly told how to solve the problem (Loibl et al., 2016; Kapur & Bielaczyc, 2012). Invention activities therefore help students to deeply understand concepts, identify the limitations of previous procedures when they do not work, and look for new patterns and interpretations that connect with and build upon their existing knowledge. Similar ideas can also be found in other "teaching for understanding" frameworks (e.g Wiske, 1997). In an assessment context, these types of activities can provide evidence about whether students can flexibly apply their knowledge schema to unfamiliar contexts.

In traditional tests, if students do not know the relevant procedure to follow there is little they can do to progress (Schwartz & Arena, 2013). Yet in the real world, we can access resources and ask more knowledgeable others for help when learning and problem solving. Assessments that challenge students to learn to solve new problems should incorporate resources for learning, because problem solving always requires some degree of knowledge. These should be carefully crafted so that they do not provide prescriptive solution steps, but rather provide opportunities to learn about core properties of the problem and encourage implementing a certain strategy that helps to make progress toward a solution. Contrasting cases represent one approach to providing such structure in learning and invention activities that has proven effective in experimental settings and that could be applied to larger-scale assessments (e.g., Shwartz & Martin, 2004; Taylor et al., 2010).

### *Design Principle 4: Provide Feedback and Instructional Support to Students During Tasks*

To complement the design principle 3, instructional support in the form of advice, feedback, or prompts as students engage in activities can promote deep learning in beginners and enable the decisions they make in their learning to be observed (Azevedo & Aleven, 2013; van Joolingen et al., 2005). Targeted feedback and scaffolded interventions can also reduce the risk that beginners disengage from an assessment because they perceive it to be beyond their capacities: it can provide clarity over task instructions, reduce the degrees of freedom or number of acts required to make progress, signal critical features that a student may have missed upon first attempt, reproduce partial solutions, and elicit further articulation or reflection questions (Guzdial et al., 2001). All these are important functions that can serve to elicit more information about students' competency level and reengage students if they appear disengaged–which is especially important in the context of extended performance tasks with less discrete items and fewer clear "data points" to inform scores.

Including feedback and instructional support in summative tests may be challenged because of fairness (e.g., unfairly penalizing students who do not need such supports) or validity concerns. However, Shute et al. (2008) evaluated the psychometric quality of an algebra assessment delivered by a digital learning system that combined adaptive task sequencing with instructional feedback: a comparison of metrics for a treated (with feedback) and control (without feedback) sample showed that providing instructional support did not make the assessment less able to detect differences between students. From a measurement perspective, the main challenge remains to develop and validate psychometric models that account for how these resources affect students' measured ability, as this ability is no longer fixed but can progress during the assessment experience (see Piacentini, 2023, and Scalise et al., 2023, for a more in-depth discussion of this issue).

### *Design Principle 5: Design Tasks That Are 'Low Floor, High Ceiling' or Otherwise Adaptive*

Next-generation assessments of 21st-century competencies should enable all students to demonstrate their ability to learn and tackle problems by using the tools and resources available to them–regardless of their initial level of knowledge or skill. Adapting assessment challenges to different abilities not only improves the quality of the measures but also the authenticity of the assessment experience: in real life, people seldom take on challenges that are too easy or impossible to achieve, yet in traditional tests this happens quite frequently. One approach to catering to different student ability levels is to design so-called "low floor, high ceiling" tasks, meaning that they are largely accessible to all ability levels while providing scope to challenge top performers. One cluster of low floor, high ceiling problems asks students to produce an original artifact: for example, a story, a game, a design for a new product, an investigative report, a speech, etc. These artifacts generate a wide range of qualitatively distinct responses and even top performers are incentivized to produce a solution that is richer, more complete, and unique. This type of design can also be used in more standardized problem-solving tasks if students are informed about intermediate targets to achieve and that progress towards more sophisticated solutions will be rewarded.

Adaptive designs can also address the complexity of measuring learning-in-action amongst heterogeneous student groups. A relatively simple way to integrate some level of adaptivity in assessment design is to structure achieving larger and more complex goals as a sequence of tasks that gradually increase in difficulty (similar to a "level-up" mechanism) and instruct students to complete as many as they can. More proficient students will quickly complete the initial set of simple tasks, after which they will encounter problems that challenge them; less prepared students are still able to engage meaningfully with the tasks and demonstrate what they can do, even if they do not complete the full sequence. Both groups of students work at the cutting edge of their abilities with obvious benefits in terms of measurement quality and test engagement. More sophisticated adaptive pathways could also be introduced: based on the quality of students' work in previous tasks, they could be directed on-the-fly towards easier or more difficult subsequent tasks.

## From principles to practice

These five design principles are intended as broad guidelines for designing next-generation assessments capable of measuring and supporting the development of 21st-century competencies. They are not intended to be prescriptive but rather illustrate the characteristics of innovative assessments that can provide valid information about students' real-life abilities. Designing assessments is a complex and challenging endeavor that must be guided by theory and research about cognition; yet just like any other design activity, scientific knowledge provides direction and constrains the set of possibilities, but it does not prescribe the exact nature of the design nor ingenuity in the final product. 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 program), the context in which it will be used (e.g., classroom or large-scale), and practical constraints (e.g., resources and time) (Pellegrino, 2023). Our fundamental argument here is not to shift completely from one assessment paradigm (i.e., using only short discrete, static items) to another (i.e., using only extended, interactive performance tasks with resources and feedback), but rather encourage the development and use of a more diversified set of assessment experiences where the breadth and depths of tasks and associated measurement models are aligned with what the assessment intends to measure.

In this section, we present two examples of innovative assessments that connect with the five design principles discussed above: one in the context of large-scale assessment–the PISA 2025 Learning in the Digital World assessment–and the other–the Platform for Innovative Learning and Assessment–in the context of formative, classroom assessment. These two examples intend to show how the principles above can be operationalized in different contexts of use.

It should also be noted that the development of valid and effective innovative assessments must follow the principles of a coherent design process. Exploiting the advantages of technology requires well-designed task models that define the features of responses that matter for scoring, as well as a clear conceptual understanding of how different patterns in student behaviors align with competency states or strategies at different levels of proficiency in the target construct. The Evidence-Centred Design (ECD) framework provides a rigorous and practical framework for making coherent design decisions, particularly in the context of measuring complex competencies in interactive environments (Mislevy & Riconscente, 2006; Piacentini, 2023).

### *Large-scale assessment: PISA 2025 Learning in the Digital World*

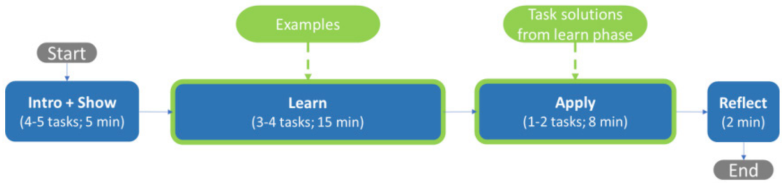
In its forthcoming 2025 cycle, the OECD's Programme for International Student Assessment (PISA) will include an assessment of Learning in the Digital World (LDW). The LDW assessment intends to provide comparable data on students' readiness to learn and problem solve using computational tools in digital learning environments. The assessment is based on a constructivist notion of learning, i.e., that learners are active participants in building their knowledge, and that knowledge and understanding are built through interactions between other people or objects. It is thus aligned with contemporary theories of cognition and aims to capture knowledge-in-use, and particularly, students' capacity to solve problems and self-regulate their learning. To distinguish more effective learners from less effective learners, the assessment provides opportunities for students to engage in knowledge construction activities. Depending on the unit context, students can use computational tools (simple programming, modelling or simulation environments) to build and represent their emergent understanding.

The assessment is structured as a learning experience (design principle 1). This diverges significantly from the traditional PISA format (a series of stimuli together with short, independent items) to a series of connected lessons within extended units of 30 minutes, organized as a series of phases (See Figure 2). First, a virtual agent introduces the overall learning goal of the unit and frames the experience as a tutoring session (e.g., "I'm going to teach you how to…"). Students then complete a pre-test battery of short, discrete items that aim to measure their prior knowledge of the concepts and practices they will learn and apply in the unit. This pre-test is intended to serve as useful information for contextualizing student behaviors and performance throughout the unit (design principle 2). The students then complete a tutorial to familiarise themselves with the core functionalities and affordances of the learning environment before spending approximately 10–15 minutes in the learning phase of the unit. Here, students complete a series of carefully scaffolded tasks that each focus on a concept or practice in the context of a simple to moderately open problem. In the final "challenge" phase of the unit, students must apply the concepts and practices they have learned throughout the unit to solve an open, complex and multi-step problem. This design aims to immerse the students in an authentic, digital learning experience following a sequence of coherent instructional tasks, starting from basic and progressing towards more complex ("low floor, high ceiling") applications (design principle 5). Two detailed prototype units are included in the assessment framework to illustrate this approach (OECD, 2023).

**Figure 2**

***Unit structure in the PISA 2025 Learning in the Digital World test***

***Source: OECD (2023).***



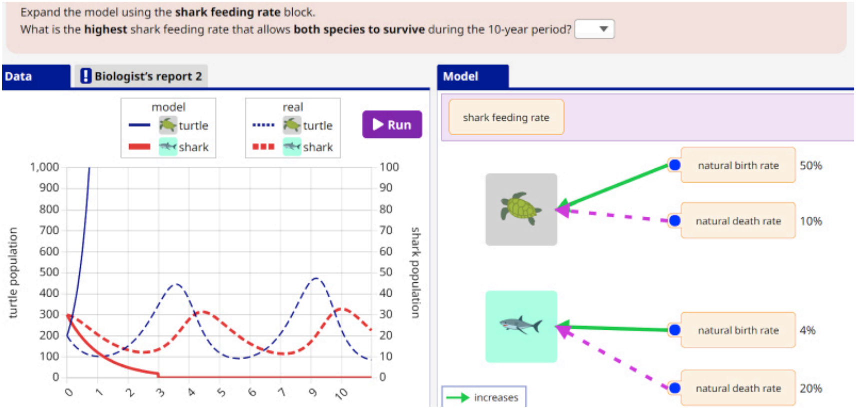

Part of the LDW assessment construct relates to students' capacity to engage in self-regulated learning (SRL), which refers to the monitoring and control of one's metacognitive, cognitive, behavioral, motivational and affective processes while learning (Panadero, 2017). From a design perspective, generating observations on SRL processes requires giving students problems that can only be solved through multiple iterations, providing them with information resources, and developing tools that students can use to monitor and evaluate their progress as well as receive feedback (design principles 3 and 4). Several affordances were embedded in the test environment for these purposes (see Roll & Barhak-Rabinowitz, 2023, for a detailed presentation). For example, students can access worked examples to help them solve a task; they can receive feedback by testing their computational artifacts (e.g., programs or models) against expected outcomes or by asking the tutor to check their work; and they can choose to view correct solutions after submitting a response to each task. At the end of each unit, students are also asked to evaluate their performance and report the effort they invested and the emotions they experienced as they worked. The assessment thus integrates the idea that we can better measure complex socio-cognitive constructs by giving students choice in the assessment, and monitoring not just how well students ultimately solve problems but also how they go about learning to do so.

One of the main design challenges for this large-scale assessment centered on how to define learning resources and feedback that could be standardized and automated, given the heterogeneous student population and limited testing technology. This was challenging for several reasons. Effective feedback is both task- and tutee-dependent: it must be based on a complete model of the task's demands, affordances, and solution space, while also adapting to the performance of the student. Without attending to both, the system cannot generate feedback that is useful for all students, in turn failing to bring each student to their zone of proximal development (Vygotsky, 1978). Artificial Intelligence holds promise for answering to this challenge, at least for some types of learning and assessment experiences (see Hu et al., 2023), but this was not practically, politically, or financially viable to integrate in the PISA 2025 cycle. In the LDW assessment, personalized feedback is provided implicitly through experimentation tools, i.e., generated by the system in response to students' actions (e.g., testing their model), or explicitly by the virtual agent; and more general feedback is available to students by accessing complete solutions upon task completion. For example, in the released unit Conservation[2], students construct a model of a marine ecosystem that can be used to predict what will happen in the future under different scenarios (See Figure 3). They are instructed that their model is accurate when its predictions (the straight lines) correspond to historical data (the dotted lines). Each time they run their model they receive visual feedback and are expected to continue improving their model parameters until they obtain a good fit with the real data.

**Figure 3**

***Excerpt from the PISA 2025 LDW released unit "Conservation"***

*Source: OECD (2023) and accessible online at* https://conservation.netlify.app.




2 Interactive example experience accessible at: https://conservation.netlify.app.

Designing effective yet standardized (i.e., non-adaptive) learning resources represented an even greater challenge. On the one hand, resources needed to be useful enough to most students so that they were incentivized to consult them and so that they provided useful hints or guidance on optimal strategies. On the other hand, resources needed to encourage the transfer of concepts without an excessive cognitive demand or without giving students prescriptive solution steps. Experts decided to address this complex trade-off by developing worked examples that students could access for each task. The worked examples provide an indication of how similar problems can be solved but students must transfer the concepts/practices described in the example to a different problem context.

The evidence needed to make claims on students' proficiency in the target knowledge, skills and attitudes of the assessment is derived from multiple sources. These include students' responses to explicit questions in the learning and challenge phase of the units, the quality and completeness of the computational artifacts (i.e., programs and models) they produce, and sequences of process data (log files) from their interactions with the environment and its affordances. The significant use of process data for scoring in the LDW assessment represents another innovation in large-scale assessment. Some uses of process data for scoring are relatively straightforward: for example, in modeling units, process data can indicate whether students completed all the steps required in constructing their model (e.g., conducting sufficient experiments to make an evidence-based conclusion on the relationship between two variables). Using process data is also essential–but much trickier–for evaluating students' self-regulated learning processes, given that any potential evidence of these behaviors must be evaluated in the context of students' previous and subsequent actions as well as the state of the learning environment when the action takes place. For example, the action of seeking help from the virtual agent may only be considered evidence of monitoring one's progress and adapting if the student actually needs help (i.e., the student is stuck) and if they implement what the tutor recommends (where this is possible). Detailed evidence rules have been developed to describe how student response and response process data should be interpreted for scoring (OECD, 2023; Foster et al., under review). These evidence rules will also be complemented by applying data mining methods to pilot and field trial data to uncover other potential evidence of coherent self-regulated learning behaviors.

Taken together, the innovations presented here for the LDW assessment align with our five design principles and respond to well-defined evidentiary needs. As further elaborated by Piacentini (2023), the assessment intends to respond to three interconnected questions: 1) what types of computational problems can students solve?; 2) to what extent are they able to learn new concepts in this domain by solving sequences of connected, scaffolded tasks?; and 3) to what extent is this learning supported by productive behaviors and learning processes, such as using resources or monitoring progress towards their learning goals? Following an ECD process, these questions have guided the domain analysis and domain modeling, defined the cognition (i.e., student) model of the assessment, oriented the design of tasks needed to elicit necessary observations about student behaviors and processes, and guided the assessment analysis plans. Following the assessment, the OECD intends to produce multi-dimensional reports of student performance including measures of their: 1) overall computational problem-solving competence (represented on a scale, as in other PISA assessments); 2) learning gains, i.e., the extent to which students' can learn and successfully apply concepts and practices during the experience; and 3) self-regulated learning behaviors. The goal is to provide policy makers with actionable information beyond a single score and position in an international ranking, including more nuanced descriptions of what students can do and cannot do, and the aspects of their performance that deserve more attention.

***Formative assessment: The Platform for Innovative Learning Assessments (PILA)***

The Platform for Innovative Learning Assessments (PILA) is a research laboratory coordinated by the OECD[3]. Its aim is to design learning and assessment experiences that provide real-time feedback on student performance, typically for use in the context of classroom instruction. One overall objective of PILA is to make assessment designers, programmers, measurement experts, and educators work together to explore new ways to close the gap between learning and assessment.

One of the assessment applications developed in PILA called Karel focuses on computational problem solving[4]. Students use a block-based visual programming

3 https://pilaproject.org/

4 Several tasks in the computational problem-solving application are publicly available as demos, accessible here: https://demo-gallery-karel.netlify.app/karel-player

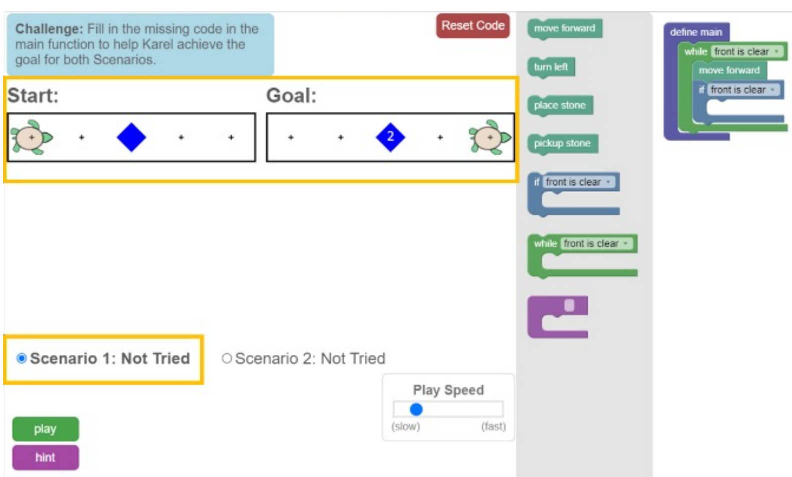
interface to instruct a turtle robot ("Karel") to perform certain actions. The tasks in the Karel application are designed to be "low floor, high ceiling" (design principle 5): low floor because the visual programming interface is intuitive enough for students who have minimal experience with programming and digital tools to still make progress, and high ceiling because the tasks are open-ended and allow for higher-performing students to challenge themselves by building "expert" solutions. programming and digital tools to still make progress, and high ceiling because the tasks are open-ended and allow for higher-performing students to challenge themselves by building "expert" solutions.

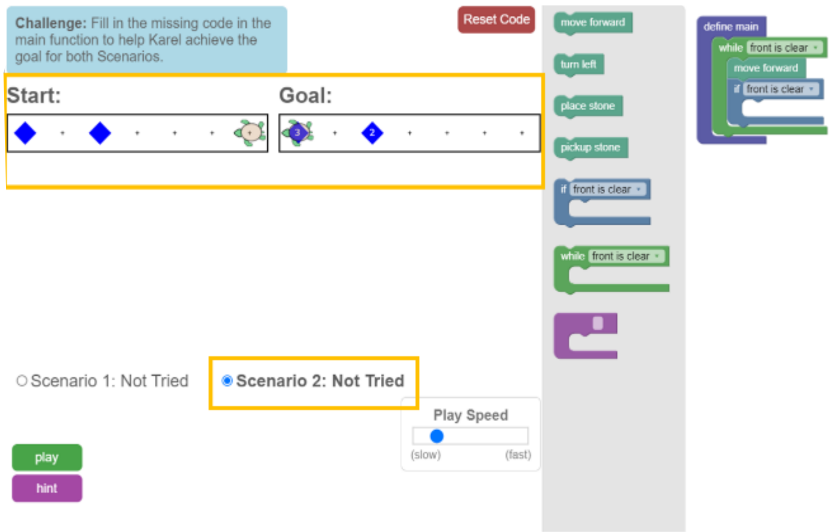
Figure 4 shows an example of a low floor, high ceiling task asking students to create a single program that moves Karel to the goal state shown in two different scenarios. To solve the problem, students can toggle between the two scenarios to visually observe the differences in the environment and how well their program works in both scenarios. Even students with solid programming skills generally require multiple iterations before finding an optimal solution for both scenarios, but the scoring models for the task take into account partial solutions (e.g., solving the problem in only one scenario).

**Figure 4**

***Example of a low floor, high ceiling task in the Karel PILA application***

Source: Platform for Innovative Learning Assessments (PILA). https://pilaproject.org/






PILA includes a series of assessment experiences designed by experts and that are structured as a progression of increasingly complex tasks with a common learning target (e.g., using functions efficiently)—thus effectively creating an extended performance task (design principle 1). As they progress through the sequence of tasks, every student will experience a point in which they struggle and need to make multiple attempts to succeed. The sequence is designed so that students are expected to transfer what they have learned in earlier tasks to more complex tasks. In the near future, PILA plans to include adaptive sequence pathways (i.e., problem sequences that adapt in real time to student performance) to further align the assessment experience with students' previous knowledge and skills (design principle 2). A related and active research line in PILA consists of developing statistical models that can accumulate information on students' current levels of knowledge and skills from all the tasks they complete in the platform. For example, when a student successfully completes a programming problem in the Karel application, the estimate of her ability in programming is updated if the problem is considered relatively difficult given her known ability level.

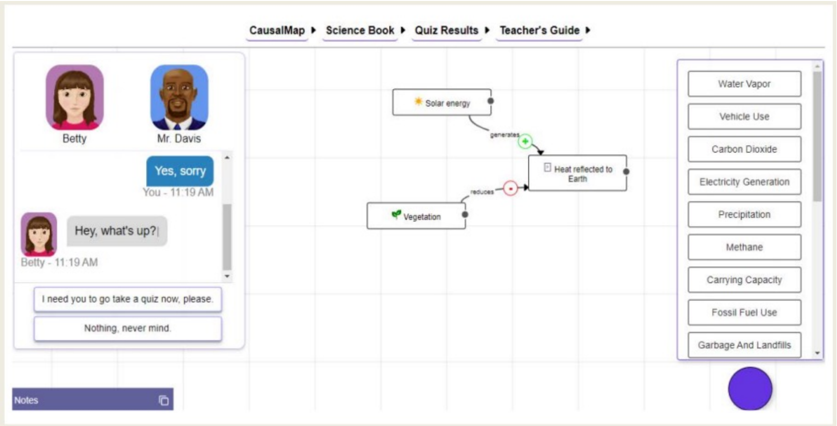
Instructional support and opportunities for students to receive feedback on their work are integrated as key features of all PILA applications (design principles 3 and 4). In the Karel application, students receive visual feedback on their program by testing and comparing the real state with the goal state. If students cannot make progress in the task they can also access learning resources, such as hints or "good" and "bad" example solutions to similar problems. In another PILA application called Betty's Brain, developed with researchers at Vanderbilt University (see Biswas et al., 2015), students engage in an extended learning-by-teaching task in which they teach a virtual agent, "Betty", about a given scientific phenomenon5. They do so by searching through hyperlinked resources in a digital textbook and constructing a concept map that represents their emergent understanding of the phenomenon[5] in question (See Figure 5 for an example task focused on climate change). Students can ask Betty to take tests using the information represented in the concept map; Betty's performance on the test then informs students if they have any wrong or missing elements in the concept map. To be successful in the environment, students must apply metacognitive strategies for setting goals, developing plans for achieving these goals, monitoring their plans as they execute them, and evaluating their progress. They have access to learning resources (the science book) and can seek feedback on their emergent knowledge and understanding (by asking Betty to take a quiz, or asking another virtual agent, "Mr Davis", for help). The data collected in this blended learning and assessment environment are thus particularly valuable for evaluating students' self-regulated learning skills.

5 An extended Betty's Brain demo task focused on learning about ecosystems is publicly available here: https://new-kl-package.bettysbrain.pages.dev/practice/resource-library

**Figure 5**

***Excerpt from the Betty's Brain PILA application***

*Source: Platform for Innovative Learning Assessments (PILA).* https://pilaproject.org/




Several other PILA applications integrating our five design principles are currently under development. Some of them focus on foundational knowledge and skills (mathematics, reading and scientific inquiry), while others target transversal skills (game design, social problem solving). Digital technologies, such as AI avatars that converse with students in natural language or tools that support collaborative problem-solving between students, are also being integrated to collect more quality evidence on complex competencies, such as collaboration and communication, in different contexts.

Assessment systems like PILA hold the promise of reducing the current separation that exists between formative and summative assessments. As students complete more and more assignments in different PILA applications, they learn important concepts and skills. At the same time, PILA collects data on what they can and cannot do in different contexts, and through a principled design we can use this evidence to construct progressively more precise inferences on their competencies across multiple domains, including difficult-to-measure competencies like self-regulation. Whether these continuous systems of assessment will eliminate the need for summative assessments is an open question that is worth exploring.

## Concluding remarks and reflections for future work

This chapter has argued that the next-generation of assessments should focus on observing and interpreting how students solve complex problems and learn to do new things. Reliance on traditional tests risks encouraging a skill-and-drill education system that does not prepare students adequately for future learning. Assessments often shape and influence the teaching and learning that takes place within education systems. If we want to shift policy into practice, we must be able to assess the competencies that students need for their future, meaning we must reproduce the features of these learning situations in assessments.

To do this, we need a good understanding of what drives productive learning and therefore what students should be expected to demonstrate in such situations. Research from the learning sciences concludes that experts (i.e., those that can use their knowledge to solve new problems) acquire competence through interactions in communities of practice, organize their knowledge into schema they continuously consolidate, and overcome the limits of their current knowledge through self-regulation. These processes can be reproduced and observed, at least to some extent, with extended assessment tasks that stimulate students to engage productively with learning resources and affordances in relatively open environments. Research on instructional design points towards some general design innovations that are worth trialing at scale in assessment, including the use of interactive and extended performance tasks that confront students with problems they have not seen before, that provide them with resources to explore and build upon their understanding, that assist them with feedback and support when they struggle to make progress, and that adapt in terms of complexity to what students can and cannot do.

Clearly, creating these types of assessment experiences is not without complexity from a design perspective; and making sure that they work from a measurement perspective is even more challenging. It will require coordinated efforts in multiple directions. First, we need a solid understanding of how learning unfolds across different domains and how to design tasks that reproduce the key elements of the social processes behind developing expertise in those domains. Second, we need improved measurement models that can validly interpret complex patterns of behaviors in dynamic environments, where both a student's knowledge and the problem state change as a result of their actions. Third, we need enhanced processes to validate the inferences we make from these more complex and performance-oriented assessments, including their sensitivity to cross-cultural differences.

The two examples presented in this chapter–the OECD's PISA LDW test and PILA platform–represent initial engagement with the enterprise of innovating assessments. Yet no single assessment can evaluate all forms of knowledge and skill that we value for students; nor can a single instrument meet all of the educational goals held by parents, practitioners and policymakers. Rather, we should aim to establish a coordinated system of assessments in which different tools are used for different purposes, but where all assessments faithfully represent the constructs of interest and model good teaching and learning practices (Pellegrino, 2023). In these systems, assessments should measure higher-order cognitive skills (i.e., those that support knowledge-in-use and 21st-century competencies) in authentic and realistic contexts, should be instructionally valuable and provide instructionally useful information about the extent to which students are capable of performing such tasks, and should be valid, reliable and fair for a range of learners (Darling-Hammond et al., 2013; Pellegrino, 2023). The major challenge is determining the conditions and resources needed to create coherent systems of assessments that work across different contexts (i.e., from classrooms to larger organizational units, such as districts, states, countries, etc.), that meet the ambitious goals we have for the educational system, that meet the information needs of different stakeholders, and that align with the criteria above.

Aspects of this assessment system design and implementation challenge are further discussed in the concluding chapter of *Innovating Assessments to Measure and Support Complex Skills* (Pellegrino et al., 2023). What is clear is that the work envisioned in that report, and summarised in this chapter, cannot be undertaken without an investment of multiple forms of capital: intellectual, financial and political. Next-generation assessment development is inherently a multidisciplinary enterprise involving different communities of experts (e.g., assessment designers, technology developers, learning scientists, domain experts, measurement experts, data scientists, educational practitioners, policy makers) that need to work together collaboratively to find solutions to the many conceptual and technical challenges that come with designing innovative assessments of complex competencies.

The development of assessments for application and use at any reasonable level of scale is a time consuming and costly enterprise, with the bulk of the substantial funds currently expended at national and international levels on assessment programs focused on traditional disciplinary domains that fall within conventional parameters for task development, delivery, data capture, scoring and reporting. Substantial fiscal capital is required to assemble and support the multidisciplinary teams needed to conduct research and development supporting the creation of innovative next-generation assessments, as well as for efforts to bring them to full maturity by scaling up their implementation when evidence exists that they can effectively address the challenge of measuring the constructs that matter. Finally, it is hard to make major changes within existing assessment systems when there are well-established operational programs that are entrenched in practice and policy. Change of the type needed requires strong political will and vision to encourage people to think beyond what is possible now or even in the near future–and that includes policy makers, but also the educational assessment development community, the measurement and psychometric community, and the educational practice community. Without political will, it will be impossible to generate sufficient fiscal capital to assemble the intellectual capital required to pursue next-generation assessment development and implementation, and in turn achieve meaningful change in educational assessment.

To illustrate these points, the development of the LDW assessment was only possible because of the convergence of these three types of capital. The innovative assessment included in each PISA cycle is now seen as a safe space to test important innovations in task design and analytical models that can be transferred to other assessments or that can provide inspiration for the development of national assessments once their value is proven. The PISA Governing Board provided the financial and political support needed to start the development of the test several years before the main data collection and to engage in extensive validation processes for its design and analytical choices. Further resources were also made available by research foundations that recognized the value of innovating assessments. The development of the assessment has also been steered by a group of experts with different disciplinary backgrounds: subject matter experts worked side-by-side with psychometricians, scholars in learning analytics and experts in UI/UX design. This cross-fertilisation was important to make space for new methods of evidence identification in digital learning environments while keeping in mind the core objective to achieve comparable metrics that result in valid interpretations of performance differences across countries and student groups.

In summary, we argue that we need many new disciplinary and cross-disciplinary assessments to provide an exhaustive description of the quality of educational experiences and the extent to which students are prepared to thrive in their futures. There is evidence that innovative assessments of educationally and socially significant competencies are both desirable and possible, and we have described some overarching characteristics of what these assessments should look like. The evidence also suggests that cooperation and collaboration on a global scale, including through the design of open-source technology and model tasks, may well be the best and only way to achieve such advances—at least for the time being.

## References

Azevedo, R., & Aleven, V. (Eds.). (2013). International handbook of metacognition and learning technologies. *Springer.* https://doi.org/10.1007/978-1-4419-5546-3

Binkley, M., Erstad, O., Herman, J., et al. (2011). Defining twenty-first century skills. In Griffin, P., McGaw, B., & Care, E. (Eds.), *Assessment and teaching of 21st-century skills.* Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Biswas, G., Segedy, J., & Bunchongchit, K. (2015). From design to implementation to practice a learning by teaching system: Betty's Brain. *International Journal of Artificial Intelligence in Education, 26*(1), 350–364. https://doi.org/10.1007/s40593-015-0057-9

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2019). Implications for educational practice of the science of learning and development. *Applied Developmental Science, 24*(2), 97–140. https://doi.org/10.1080/10888691.2018.1537791

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

DiCerbo, K., Shute, V., & Kim, Y. (2017). The future of assessment in technology-rich environments: Psychometric considerations. In Spector, J., Lockee, B., & Childress, M. (Eds.), *Learning, design, and technology: An international compendium of theory, research, practice, and policy.* Springer International Publishing. https://doi.org/10.1007/978-3-319-17727-4_66-1

Dumont, H., Istance, D., & Benavides, F. (Eds.). (2010). *The nature of learning: Using research to inspire practice.* OECD Publishing. https://doi.org/10.1787/9789264086487-en

Ericsson, K. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In Ericsson, K., Charness, N., Feltovich, P., & Hoffman, R. (Eds.), *The Cambridge handbook of expertise and expert performance.* Cambridge University Press. https://doi.org/10.1017/cbo9780511816796.038

European Commission (2019). *Key competences for lifelong learning*. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/569540

Fadel, C., & J. Groff (2018). Four-dimensional education for sustainable societies. In Cook, J. (Ed.). *Sustainability, human well-being, and the future of education.* Springer International Publishing. https://doi.org/10.1007/978-3-319-78580-6_8

Foster, N., Holmes, J., Linsenmayer, E., Zoanetti, N., & Yildiz, H. (under review). Developing rule-based scoring methods that capture student progress in computational problem solving in open, interactive tasks. *Learning and Individual Differences.*

Foster, N. (2023). 21st-century competencies: Challenges in education and assessment. In Foster, N., & Piacentini, M. (Eds.), I*nnovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Foster, N., & Piacentini, M. (Eds.). (2023). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Guzdial, M., Rick, J., & Kehoe, C. (2001). Beyond adoption to invention: Teacher-created collaborative activities in higher education. *Journal of the Learning Sciences, 10*(3), 265–279. https://doi.org/10.1207/s15327809jls1003_2

Hatano, G. (1990). The nature of everyday science: A brief introduction. *British Journal of Developmental Psychology, 8*(3), 245–250. https://doi.org/10.1111/j.2044-835x.1990.tb00839.x

Hatano, G., & Oura, Y. (2003). Commentary: Reconceptualizing school learning using insight from expertise research. *Educational Researcher, 32*(8), 26–29. https://doi.org/10.3102/0013189x032008026

Hu, X., Shubeck, K., & Sabatini, J. (2023). Artificial Intelligence-enabled adaptive assessments with Intelligent Tutors. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences, 21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Kyllonen, P. C., & Sevak, A. (2024). *Charting the future of assessments.* ETS Research Institute. https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

Loibl, K., Roll, I., & Rummel, N. (2016). Towards a theory of when and how problem solving followed by instruction supports learning. E*ducational Psychology Review, 29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Marion, S. F., & Pellegrino, J. W. (2007). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57. https://doi.org/10.1111/j.1745-3992.2006.00078.x

Mislevy, R. (2018). *Sociocognitive foundations of educational measurement.* Routledge.

Mislevy, R., & Riconscente, M. (2006). Evidence-centered assessment design. In Downing, S., & Haladyna, T. (Eds.), *Handbook of test development.* Lawrence Erlbaum.

Nunes, C., Nunes, M., & Davis, C. (2003). Assessing the inaccessible: Metacognition and attitudes. *Assessment in Education: Principles, Policy & Practice, 10*(3), 375–388. https://doi.org/10.1080/0969594032000148109

OECD (2023), *PISA 2025 Learning in the Digital World assessment framework (second draft).* https://www.oecd.org/content/dam/oecd/en/topics/policy-sub-issues/learning-in-the-digital-world/

OECD (2018). *The future of education and skills 2030.* OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2018/06/the-future-of-education-and-skills_5424dd26/54ac7020-en.pdf

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology,* 8, https://doi.org/10.3389/fpsyg.2017.00422

Pellegrino, J. (2023). Introduction: Arguments in favour of innovating assessments. In Foster, N., & Piacentini, M. (Eds.). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Pellegrino, J., Chudkowsky, N., & Glaser, R. (Eds.). (2001). K*nowing what students know: The science and design of educational assessment.* National Academy Press. http://faculty.wiu.edu/JR-Olsen/wiu/common-core/precursor-documents/KnowingWhatStudentsKnow.pdf

Pellegrino, J., DiBello, L., & Goldman, S. (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., Foster, N., & Piacentini, M. (2023). Conclusions and implications. In Foster, N., & Piacentini, M. (Eds.). *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Pellegrino, J., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st-century.* National Academies Press. https://doi.org/10.17226/13398

Piacentini, M. (2023). Defining the conceptual assessment framework for complex competencies. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Piacentini, M., Foster, N., & Nunes, C. (2023). Next-generation assessments of 21st-century competencies: Insights from the learning sciences. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills. OECD Publishing.* https://doi.org/10.1787/e5f3e341-en

Roll, I., & Barhak-Rabinowitz, M. (2023). Measuring self-regulated learning using feedback and resources. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Roll, I., Aleven, V., McClaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267–280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Roll, I., Holmes, N. G., Day, J., & Bonn, D. (2012). Evaluating metacognitive scaffolding in guided invention activities. I*nstructional Science, 40*(4), 691–710. https://doi.org/10.1007/s11251-012-9208-7

Sabatini, J., Hu, X., Piacentini, M., & Foster, N. (2023). Designing innovative tasks and test environments. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills. OECD Publishing.* https://doi.org/10.1787/e5f3e341-en

Schwartz, D., & Arena, D. (2013). *Measuring what matters most: Choice-based assessments for the digital age.* The MIT Press. https://doi.org/10.7551/mitpress/9430.001.0001

Schwartz, D., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics education. *Cognition and Instruction, 22*(2), 129–184. https://doi.org/10.1207/s1532690xci2202_1

Simon, H. (1980). Problem solving and education. In Tuma, D., & Reif, R. (Eds.), *Problem solving and education: Issues in teaching and research.* Erlbaum.

Scalise, K., Malcolm, C., & Kaylor, E. (2023). Analysing and integrating new sources of data reliably in innovative assessments. In Foster, N., & Piacentini, M. (Eds.), *Innovating assessments to measure and support complex skills.* OECD Publishing. https://doi.org/10.1787/e5f3e341-en

Scott, C. (2015). The futures of learning 2: What kind of learning for the 21st-century? Education, Research and Foresight: Working Papers. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000242996

Shute, V., Hansen, E., & Almond. R. (2008). You can't fatten a hog by weighing it—or can you? Evaluating an assessment for learning system called ACED. I*nternational Journal of Artificial Intelligence in Education, 18*(4), 289–316.

Shute, V., & Kim, Y. (2013). Formative and stealth assessment. In Spector, J., et al. (Eds.), *Handbook of Research on Educational Communications and Technology (4th Edition).* Lawrence Erlbaum.

Taylor, J. L., Smith, K. M., van Stolk, A. P., & Spiegelman, G. B. (2010). Using invention to change how students tackle problems. *CBE—Life Sciences Education, 9*(4), 504–512. https://doi.org/10.1187/cbe.10-02-0012

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal, 42*(3), 454–476. https://doi.org/10.1002/berj.3215

Thornton, S. (2012). Issues and controversies associated with the use of new technologies. In Gormley-Heenan, C., & Lightfoot, S. (Eds.), *Teaching Politics and International Relations.* Palgrave Macmillan UK. https://doi.org/10.1057/9781137003560_8

van Joolingen, W., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior, 21*(4), 671–688. https://doi.org/10.1016/j.chb.2004.10.039

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes.* Harvard University Press.

Wiske, M. (Ed.). (1997). *Teaching for understanding: Linking research with practice.* Jossey-Bass.

World Economic Forum (2015). New vision for education: Unlocking the potential of technology. https://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658