

EXTENDED EXECUTIVE SUMMARY

# INNOVATING ASSESSMENTS TO MEASURE AND SUPPORT COMPLEX SKILLS



This work is published under the responsibility of the Secretary-General of the OECD. The opinions expressed and arguments employed herein do not necessarily reflect the official views of OECD member countries.

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

© OECD 2023

Photo credits: Cover design on the basis of images from © Shutterstock/treety; © Shutterstock/Merfin.

You can copy, download or print OECD content for your own use, and you can include excerpts from OECD publications, databases and multimedia products in your own documents, presentations, blogs, websites and teaching materials, provided that suitable acknowledgment of OECD as source and copyright owner is given. All requests for commercial use and translation rights should be submitted to [rights@oecd.org](mailto:rights@oecd.org).

## ABOUT THIS PUBLICATION

This brochure summarises the key messages of the OECD publication *Innovating Assessments to Measure and Support Complex Skills* (Foster and Piacentini, 2023). *Innovating Assessments* is the product of a collaborative effort between the OECD Secretariat and the PISA Research and Innovation Group (RIG), as well as several other international experts and collaborators in the field of educational measurement and assessment design.

Both the *Innovating Assessments* publication and this brochure were made possible with the generous support of Instituto Unibanco, namely by Ricardo Henriques, João Marcelo Borges and through the collaboration of the team including Djana Contier Fares, Carolina Fernandes, Valquiria A. N. Parlagreco, and Tatiana F. Laganá who provided critical review of this brochure as a consultant.

Natalie Foster and Mario Piacentini edited the original volume and contributed several chapters. RIG members, including Kadriye Ercikan, Xiangen Hu, Cesar A. Amaral Nunes, James Pellegrino, Ido Roll and Kathleen Scalise, and invited collaborators, including Miri Barhak-Rabinowitz, Hongwen Guo, Han Hui Por, Errol Kaylor, Cassie Malcom, Argenta Price, John. P. Sabatini, Keith Shubeck and Carl Wieman, contributed the remaining chapters and provided expert advice and feedback on the overall publication. Andreas Schleicher, OECD Director for Education and Skills, and Yuri Belfali, Head of the Early Childhood and Schools Division at the OECD, provided additional guidance and feedback. This brochure was prepared by Mario Piacentini, Natalie Foster and Marc Fuster (OECD).

---

Foster, N. and M. Piacentini (eds.) (2023), *Innovating Assessments to Measure and Support Complex Skills – Extended executive summary*, OECD Publishing, Paris, <https://www.oecd.org/pisa/innovation>.

# TABLE OF CONTENTS

<b>EDITORIAL</b>	<b>05</b>
<b>THE CASE FOR INNOVATING ASSESSMENTS</b>	<b>08</b>
Assessment matters	08
Shifting education goals: A focus on 21st century competencies	09
Assessing 21st century competencies calls for innovating assessment design	11
<b>NEXT-GENERATION ASSESSMENTS: DESIGN PRINCIPLES AND EXAMPLES</b>	<b>13</b>
Assessment as a process of reasoning from evidence	13
Innovating the <i>cognition</i> vertex: Defining assessment constructs	17
Innovating the <i>observation</i> vertex: Including more varied and interactive assessment tasks	28
Innovating the <i>interpretation</i> vertex: Making sense of assessment observations	41
<b>INNOVATING ASSESSMENTS: THE ROAD AHEAD</b>	<b>49</b>
Investing in next-generation assessments	49
International large-scale assessments: Possibilities for innovation at scale	53
Coda: Returning to the three types of capital	56
<b>REFERENCES</b>	<b>57</b>

# EDITORIAL

More than 20 years on from its first cycle, PISA (the Programme for International Student Assessment) has become an established and influential force for education reform. The transformational idea behind PISA lay in testing the skills of students directly through an international metric; linking that with data from students, teachers, schools and systems to understand performance differences; and harnessing the power of international collaboration to act on the data.

From its inception, PISA differed from traditional assessments. To do well in PISA, students had to be able to extrapolate from what they know, think across the boundaries of subject-matter disciplines, and apply their knowledge creatively in novel situations – rather than mainly reproduce knowledge they had learnt in class. The modern world no longer rewards us for what we know, but for what we can do with what we know. As content becomes increasingly accessible, and more routine cognitive tasks become digitised and outsourced, the focus must shift to enabling people to become lifelong learners. Epistemic knowledge – thinking like a scientist or mathematician – and ways of working are taking precedence over knowing specific formulae, names or places.

This vision of education is reflected in many contemporary frameworks calling for the development of so-called 21<sup>st</sup> century skills – including the OECD’s Learning Compass 2030. Yet without substantial changes in our education systems, the gap between what they provide our young people with and what our societies demand is likely to widen further.

One integral component of education systems is assessment. The way students are tested has a big influence on the future of education, because it signals the priorities for the curriculum and instruction. Tests will always focus our thinking about what is important, and so they should – teachers and school administrators, as well as students, will pay attention to what is tested and adapt accordingly. A fundamental question is how we can get assessment right and ensure that it helps teachers and policy makers track progress in education in ways that matter.

The trouble is that many assessment systems are poorly aligned with the curriculum and with the knowledge and skills that young people need to thrive. When designing assessments, we often trade gains in validity and relevancy for gains in efficiency and reliability.

But these priorities have a price: the most reliable and efficient test is one where students respond in ways that allow for little ambiguity – typically a multiple-choice format. A relevant test is one where we test for a wide range of knowledge and skills considered important for success in life and work.

To do this well requires multiple response formats, including open formats, which elicit more complex responses. Necessarily, these require more sophisticated marking processes. Good tests should also provide a window into students' thinking and understanding, revealing the strategies a student uses to solve a problem and providing productive feedback, at appropriate levels of detail, to fuel improvement decisions. Digital assessments, by logging traces of students' actions and not just their responses, provide several opportunities to advance assessment along these lines.

Beyond that, assessments need to be fair and ensure adequate measurement at different levels of detail so they can serve decision-making needs at different levels of the education system. We also need to work harder to bridge the gap between summative and formative assessments. The origins of education were in apprenticeship, where students learned from and with people, with immediate and personal feedback on their progress. Centuries later, the industrialisation of education then divorced learning from assessment, asking students to pile up years of learning and then calling them back much later to reproduce what they learned in often narrow and time-constrained settings. This has contributed to learning and teaching that is often shallow and focused on what can be easily measured. Digitalisation provides us now with the opportunity to re-integrate learning and assessment, to combine summative and formative elements of assessment, and to create coherent multi-layered assessment systems that extend from students to classrooms to schools to regional, national and even international levels. Better integrating assessment and learning will mean that teachers no longer see testing as taking away valuable time from learning, but rather an instrument that adds to it.

Of course, all of this also applies to PISA. PISA is viewed as an important measure of the success of school systems around the world and, as such, needs to lead education reform. Since 2012, and thanks to the introduction of computer-based delivery, PISA has expanded its range of metrics to include a new interdisciplinary domain in every cycle – including problem solving (2012), collaborative problem solving (2015), global competence (2018) and, most recently, creative thinking (2022).

In 2020, PISA went a step further: despite the most challenging of global circumstances, countries decided to invest more resources in developing innovative assessments, establishing a new Research, Development and Innovation (RDI) programme led by a group of international senior experts in assessment.

In some ways, this publication was borne out of our collaboration with different experts over the last three years since our ongoing research programme began. It makes the case for why we need to innovate assessments, explains what we need to change and how we can leverage technology in order to get there. It also makes clear that this change will not happen overnight: there is much work yet

to be done, and it will require the convergence of political, financial and intellectual capitals to bring these ideas to scale.

PISA can become an engine to drive this change forward, by harnessing the power of international collaboration between educators, researchers and policymakers, and sharing the costs – both financial and political – among countries in the search for innovative practices. Research and innovation in large-scale assessment has always been a core part of PISA’s DNA, and we are committed to continue as a global leader on the path ahead.

**Andreas Schleicher**

Director for Education and Skills  
Special Advisor on Education Policy to the Secretary-General

# THE CASE FOR INNOVATING ASSESSMENTS

This brochure summarises the key messages of the OECD publication *Innovating Assessments* (Foster and Piacentini, 2023), the product of a collaborative and multi-year research effort between international experts in the field of educational measurement and assessment and the OECD Secretariat.

The reason for engaging in this work – indeed, the case for innovating assessments – is driven by a set of interrelated propositions. The first one is that we should care about assessment. Educational assessments are important signposts indicating what students should learn and what they can do. As such, they are intricately linked to curricula and pedagogies, driving or holding back changes in educational goals and practices. The second proposition follows from the first: educational assessment should focus on *what matters*. What is worth knowing, doing, and being is subject to constant debate, with a global narrative calling to rethink what is being taught and learnt at school to better prepare students as citizens and future professionals. Connecting these two propositions is the idea that any discussion on the need to equip individuals with so-called ‘21st century competencies’ should also be a discussion on assessment. That said, shifting the focus of assessments to ‘what matters’ will only be valuable insofar as assessments are capable of measuring what they claim to measure. The third proposition is thus that assessments should measure what matters and they should measure it well.

## ASSESSMENT MATTERS

Teachers, students, and local and national policy makers often take their cues about the goals for instruction and learning from the types of tasks found on local, national, and international assessments. Assessments signal to multiple audiences what knowledge, skills, and abilities matter and illustrate the types of performance we want students to be capable of exhibiting. Thus, what we choose to assess in areas such as science, mathematics, literacy, 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, as well as the kinds of learning, we want to emphasise in our classrooms such that they can function positively within the education system.



From the system perspective, there is little point in investing heavily in curriculum and educator training reform without also investing in assessment. Curricula, pedagogy, and assessment are intricately linked and should be aligned in well-functioning education systems. Shifts in curricula and pedagogy can be driven by changes in assessment focus and by the educational gaps that they reveal, in turn informing policymaking and reform. Focusing on assessment brings clarity on teaching and learning expectations at different educational levels, contributing to establish a shared understanding of what matters and how it should be taught. The key question then becomes: exactly what matters?

## **SHIFTING EDUCATION GOALS: A FOCUS ON 21ST CENTURY COMPETENCIES**

For over 20 years now, a growing number of business leaders, educational organisations, and researchers have begun to call for new education policies that target the development of broad, transferable skills and knowledge, often referred to as “21st century skills” (e.g. see Pellegrino and Hilton, 2012; Bellanca, 2014). Such calls are grounded on the idea that success in global contemporary society and in a changing world of work demands a wider set of capabilities that go beyond the traditional literacies of reading, mathematics and science.

This rhetoric essentially argues that education should focus on the capacity to process (new) information and solve problems, which includes equipping individuals with strong disciplinary knowledge but also analytical, creative, and critical thinking skills. It should focus on broader abilities related to oneself and others too, such as social and emotional skills, tolerance, and mutual respect, and the capacity to self-regulate and better understand one’s own thinking and learning processes.

Certainly, these capabilities have always been important. Yet, in a world where work was defined by manual and routine tasks, and where the instant communication and information technologies of today were only a product of imagination, only some individuals were expected to develop them. In today’s knowledge economies, characterised by more dynamic and multicultural structures where citizens communicate instantly and self-organise, both locally and globally, advanced cognitive and socio-cognitive competencies are expected as the norm.

### **UNDERSTANDING 21ST CENTURY COMPETENCIES**

Starting before the turn of the century, a growing body of research has examined this global narrative, producing a variety of international frameworks that describe the knowledge, skills and attitudes that young people need for the future. There is a diversity of terminologies employed interchangeably within this relatively crowded space: ‘21st century skills/competencies’, ‘soft skills’, ‘interdisciplinary skills’ and ‘transferable skills’, to name just a few. This terminological ambiguity extends to the ways in which different

frameworks define specific competencies (e.g., ICT literacy vs. digital literacy vs. media literacy).

For the sake of clarity, *Innovating Assessments* uses the term ‘21st century competencies’ to refer to the broad vision of education set forth by these frameworks and to the various competencies that they describe. Although frameworks vary, they tend to describe 21st century competencies as being:

- transversal (i.e. relevant or applicable in many fields);
- multidimensional (i.e. encompassing knowledge, skills and attitudes); and
- associated with higher-order skills and behaviours that represent the ability to transfer knowledge, cope with complex problems and adapt to unpredictable situations (Voogt and Roblin, 2012).

Beyond general convergence around these core characteristics, frameworks identify, organise and classify 21st century competencies in different ways. Some group competencies based on their conceptual features, for example, cognitive, interpersonal and intrapersonal competencies (Pellegrino and Hilton, 2012). Others group competencies according to their purpose or context of use, such as ‘ways of thinking’, ‘ways of living in the world’, ‘ways of working’ and ‘tools for working’ (Binkley et al., 2012).

Abstracting from the specificities of each framework, some broadly distinct categories of competencies do consistently emerge (see Figure 1). In general, some combination of these six categories captures the essence of the exhaustive lists of competencies identified across different frameworks, with critical thinking, creative thinking, communication, and ICT-related competencies, as well as the civics and citizenship dimension, consistently appearing. Note, however, that not all frameworks include each of the broad categories identified below, nor do they always assign specific competencies to the same broader categories.

**Figure 1.** Broad categories of 21<sup>st</sup> century competencies



**Source:** Foster (2023), chapter 1 in *Innovating Assessments*.

Identifying common categories of 21st century competencies provides some useful insight about the ways in which the broader goals of education are changing. Nevertheless, these competencies are complex constructs, and eliciting valid evidence and interpretations of what students think and can do when engaging them poses several challenges. Assessing 21st century competencies well requires innovating assessment design and experiences, all the way from defining assessment constructs to designing assessment tasks and finding the right methods to interpret the evidence emerging from them.

## **ASSESSING 21ST CENTURY COMPETENCIES CALLS FOR INNOVATING ASSESSMENT DESIGN**

The first issue when assessing 21st century competencies relates to defining what to assess. These competencies are complex; they involve multiple components that are strongly intertwined in practice. On the one hand, engaging them entails activating a combination of knowledge, skills and attitudes – for instance, the ability to communicate effectively involves some language knowledge, a degree of written, verbal, or digital skill, and certain attitudes towards those with whom one is communicating. These constituent elements can also be different in different contexts of practice. On the other hand, engaging one ‘type’ of competency in real life often requires engaging other ‘types’ simultaneously. Successful problem solving, for instance, involves aspects of metacognition and self-regulation and, depending on the context and typology of the problem, it could involve creative thinking and collaboration. These complex links make it difficult to break down constructs into discrete and independently measurable components, as well as isolate and attribute evidence generated by students to one particular competency or another.

In parallel, 21st century competencies are defined, at least in part, by thought processes and behaviours that go beyond the capacity to reproduce content knowledge. For instance, the ability to critically appraise unfamiliar pieces of information depends on being able to understand what additional information needs to be searched for and how, to plan for and execute a strategy to do so, and to persist in solving the task and/or decide when to call on for help or feedback. These behaviours and ways of thinking need to be made visible in assessments for any claim on student competence to be made. For many 21st century competencies, this means designing assessment environments that provide students with tools for doing and making, and with choices and opportunities to explore and iterate upon their ideas. These kinds of affordances call for moving assessment tasks and features beyond the static, closed-response item types typically used in large-scale assessment, to generate a richer set of data on how students think and act.

Creating the next generation of educational assessments that respond to this vision of 21st century education therefore presents a sequence of challenges that assessment designers must overcome, including being able to define the target constructs of assessment, identify the relevant situations where these can be observed, replicating their core features in assessment environments, translating traces of actions within these environments into evidence, and

developing suitable models to interpret and score the evidence to make robust claims on performance.

Drawing on the key messages and most advanced examples of practice included in the OECD's *Innovating Assessments* publication, the following sections shed light on the path forward for assessment designers - including unpacking the key decisions they need to consider and the emerging tools that can help them along the way. Closing this document are some considerations on the role that education authorities can play, together with other stakeholders, as part of a broader framework of international collaboration to move the 'Next-Generation Assessments' agenda forward.

# NEXT-GENERATION ASSESSMENTS: DESIGN PRINCIPLES AND EXAMPLES

Assessing educational outcomes is not as straightforward as measuring height or weight. Assessments do not offer a direct pipeline into a student's mind; the attributes to be measured are mental 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 and can do. Deciding what to assess and how to do so is not as simple as it might appear. This is even more the case when the targets of assessment are complex constructs and performances.

*Innovating Assessments* presents key ideas on how to design the next generation of assessments that measure the competencies students need and provides actionable information to assessment designers, educators and policymakers. Measuring *what matters* calls for innovating all phases of assessment design – from what we assess to how we do it. Measuring *what matters well* entails doing so through a principled design process and leveraging digital technologies to generate relevant evidence about students' competencies and to apply innovative analytical methods for making sense of such evidence.

## ASSESSMENT AS A PROCESS OF REASONING FROM EVIDENCE

The process of making 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 achievement tests, to computerised tutoring programs, to the conversations students have with their 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 do not provide their own meaning, their value as evidence can arise only through some interpretational framework. Educational assessments provide data such as written essays, marks on answer sheets, presentations of projects, or students' explanations of their problem solutions, but these data become evidence only with respect to conjectures about how students acquire knowledge and skill.

Pellegrino and colleagues (2001) portray this process of reasoning from evidence as a triad of three interconnected elements: the

Assessment Triangle (see Figure 2). The vertices of the Triangle represent the three key elements underlying any 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 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. 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. Pellegrino, et al., 2016).

**Figure 2.** The Assessment Triangle

**COGNITION**

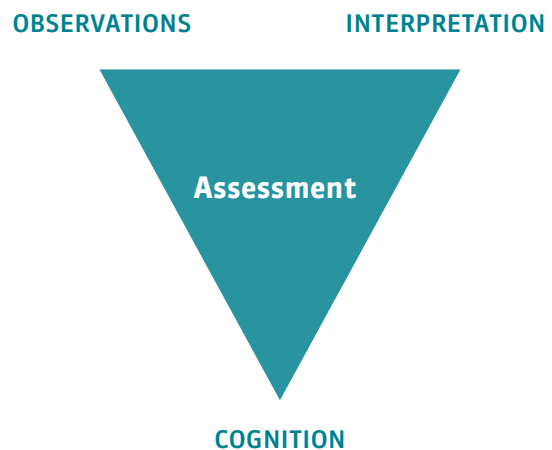
Theories, models & data about how students represent knowledge & develop competence in a domain of instruction and learning.

**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)

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; thermodynamics). For any assessment, a theory of competence in the domain 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 learning. A central premise is that the cognitive theory should represent the most scientifically credible understanding of the typical ways in which learners represent knowledge and develop expertise in the focus 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 must be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of inferences and decisions that will be made on the basis of 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.

Assessments also require 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 one or more tasks, and what the performances mean about the student's knowledge and understanding.

A crucial point is that each of the three elements of the Assessment Triangle must not only 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. Recognising that assessment is an evidentiary reasoning process, it has proven useful to be systematic in framing the process of assessment design as an Evidence Centered Design process (e.g. Mislevy and Haertel, 2006; Mislevy and Riconscente, 2006) – see Figure 3 for an overview of the different components of the ECD model.

**Figure 3.** Assessment design as an Evidence Centred Design process

Phases of defining the conceptual framework of an assessment

**DEFINING  
OBJECTIVES AND  
FOCUS**

**DEFINING THE DOMAIN OF THE ASSESSMENT**

- **Gathering information about the domain** (domain analysis), including its main components and the range of problems and situations in which people use the target knowledge and skills.
- Domain modelling: **Specifying assessment claims** (what we wish to measure), data (how we are going to measure it) and warrants (explaining why the measurement approach is appropriate).

**DEFINING WHAT STUDENT PERFORMANCE IN THE DOMAIN LOOKS LIKE (THE STUDENT MODEL)**

- **Defining the variables** (knowledge, skills and attitudes) we want to make claims on, the relationships between these variables, and whether these variables are dynamic (if some learning is expected).
- Providing a detailed vision of what students understand and can do **at different levels of proficiency**, from lowest to higher levels of mastery in each variable.

**DEFINING THE SITUATIONS WHERE EVIDENCE OF PERFORMANCE CAN BE FOUND (THE TASK MODEL)**

- **Specifying the tasks** where test-takers can demonstrate proficiency, such as in pre-defined questions or tasks (e.g., multiple-choice items, reordering or completion tasks) or in environments where the situation is shaped by test-takers' actions (e.g., simulations, games).
- **Defining drivers of complexity** and knowledge involved, **and** the **resources** embedded in the task including feedback or scaffolds to facilitate learning (if learning is expected).

**DEFINING PERFORMANCE SCORES AND INDICATORS (THE EVIDENCE MODEL)**

- Defining the evidence rules: **associating a score or value to what test-takers do** (e.g., answering questions correctly/incorrectly, taking certain actions/decisions in a given situation).
- Building a statistical model that **summarises data across tasks** in terms of updated beliefs about student-model variables.

**Source:** Piacentini (2023), Chapter 6 in *Innovating Assessments*.

**OPERATIONALISING  
THE ASSESSMENT  
(ECD FRAMEWORK)**



## **INNOVATING THE COGNITION VERTEX: DEFINING ASSESSMENT CONSTRUCTS**

In assessment development, no other issue is as critical as clearly delineating the target domain and describing the constituent knowledge, skills, attitudes, and contexts of application that underpin performance in that domain. Indeed, if the domain is ill-defined, no amount of care taken with other test development activities nor complex psychometric analysis once data have been collected will compensate for this inadequacy (Mislevy and Riconscente, 2006). It is far more likely that an assessment achieves its intended purpose when the nature of the construct guides the design of relevant tasks as well as the development of construct-based scoring criteria and rubrics (Messick, 1994).

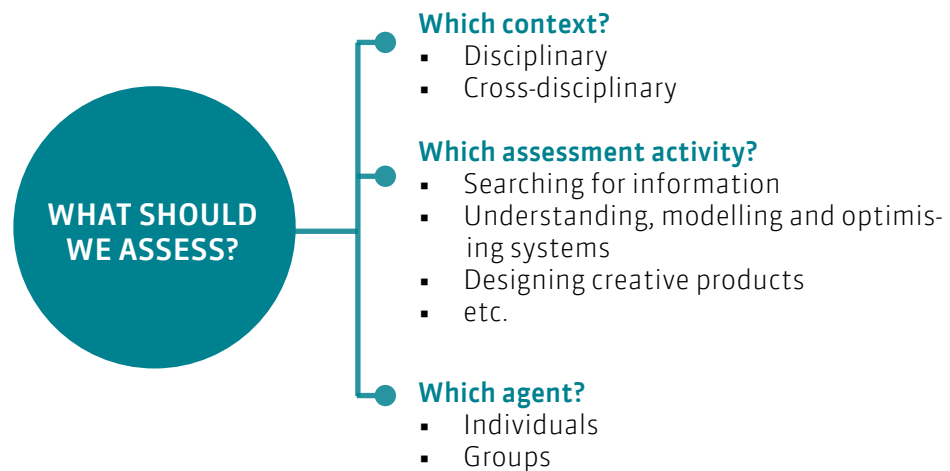
As already discussed, this critical activity becomes more challenging as the complexity of the domain and target construct(s) increases. The types of problems or learning activities that require and engage 21st century competencies call upon a different combination of knowledge, skills and attitudes, and the context of application clearly matters too for determining which of those elements are most important and how exactly they might be expressed. This means that it is important to be explicit from the initial stages of assessment design about what we expect students to demonstrate through their performance on the test.

### **EARLY DESIGN DECISIONS ON THE FOCUS OF ASSESSMENTS OF 21ST CENTURY COMPETENCIES**

When it comes to deciding what to assess, picking one framework or list of 21st century competencies and creating a single assessment instrument for each competency described might not be the best way forward. Because 21st century competencies are multi-dimensional and strongly interconnected in practice, a more productive strategy may consist of developing assessments of how students create knowledge and solve different types of complex problems, on their own or collaboratively, in different contexts of application. Making sense of what students do in open, extended problem-solving activities can give us information on how they can mobilise multiple 21st century competencies in more authentic scenarios.

As reflected in Figure 4, three interrelated questions may offer particularly useful guidance for determining the focus of the next generation of assessments:

**Figure 4.** Early-design decisions on the focus of 21st century assessments



**Source:** Piacentini and Foster (2023), Chapter 3 in *Innovating Assessments*.

- **For which kinds of performances and related activities am I interested in understanding students' preparedness?** This decision relates to explicitly defining assessment activities and the relevant practices we want students to demonstrate while engaging in those activities.
- **In which contexts of practice can students engage in assessment activities?** This decision relates to acknowledging the knowledge, skills and attitudes that students need in a given type of activity in a given context of practice (i.e. situating the activity within the boundaries of a discipline, or making it cross-disciplinary and specifying the context of application).
- **Will the assessment be organised as an individual or a group activity?** This decision relates to explicitly defining whether, when and for what purposes an assessment may provide students with the possibility of interacting with other agents, real or simulated.

### **Relevant activities for assessing 21st century competencies**

Solving complex problems engages a variety of cognitive, metacognitive, attitudinal, and socio-emotional skills. However, not all assessment problems can give us such a rich body of evidence on learners. Traditional models of problem solving, known as phase models (e.g. Bransford and Stein, 1984), suggest that all problems can be solved if we: (1) identify the problem; (2) generate alternative solutions; (3) evaluate those solutions; (4) implement the chosen solution; and (5) evaluate the effectiveness of the solution. While these descriptions of general processes are useful, they might wrongly suggest that problem solving is a uniform activity (Jonassen, 1992). In reality,

problems vary in many important ways, including the context in which the problems occur, their level of structure or openness, and the combination of skills that the problem solver has to use in order to reach a solution.

There are a variety of problem goals and activities that can be presented to students to assess 21st century competencies. For example, clusters of assessment activities that are likely to provide valid evidence on whether learning experiences have prepared students for their future include: (1) searching for, evaluating, and sharing information; (2) understanding, modelling and optimising systems; and (3) designing creative products. This is not an exhaustive typology of assessment activities; the types of problems and activities that students will need to be prepared for continues to evolve. Moreover, these three clusters of activities are not mutually exclusive and overlap to some extent. However, they are illustrative of problem types that draw distinctly different sets of competencies and related knowledge, skills and attitudes. Box 1 provides some examples of what next-generation assessments of the first cluster of activities could look like.

#### BOX 1.

### RELEVANT ACTIVITIES FOR ASSESSMENTS OF 21ST CENTURY COMPETENCIES

#### Searching for, evaluating, and sharing information

In this class of activities, the main problem-solving or learning goal consists of searching and using information to make an argued conclusion. The sequencing of tasks in an assessment should stimulate students to identify their information needs, locate information sources in online or offline environments, extract, organise and compare information from each source, reconcile conflicts in information, and take decisions on what information to share and how. This set of activities is frequently defined as information problem solving (Brand-Gruwel et al., 2005, Wolf et al., 2003). Research shows that many students are not able to solve information problems successfully (Bilal, 2000; Large and Beheshti, 2000). These activities focus on how students interact with various types of media and can be applied to virtually any area of knowledge (i.e. context of practice). They emphasise critical thinking, synthesis and argumentation, responsible communication, and self-regulated learning skills as core competencies.

There are several examples of assessments that focus on information problems. In some cases, the assessment is fully integrated in a learning experience, and the evidence is extracted in a 'stealth way' by analysing the sequences of choices students make and the result of their information search. For example, in the Betty Brain environment (Biswas, 2015) students teach a virtual agent, Betty, about a scientific phenomenon. They do so by searching through hyperlinked resources and constructing a concept map that represents their emergent understanding of the phenomenon. Students can ask Betty to take tests where she responds using the information represented in the concept map; Betty's performance on this test informs students about wrong or missing elements in the map.





Other examples embed information search and management tools within virtual worlds. The NAEP SAIL Virtual World for Online Inquiry project (Coiro et al., 2019) developed a virtual platform simulating a micro-city, where students are presented with an open learning challenge (e.g. to find out whether an historical artefact should be displayed in the local museum) and they build their knowledge by planning an inquiry strategy with a virtual partner, asking questions to virtual experts, searching for information on a web environment or in a virtual library, and using different digital tools to take notes and redact a report. The environment includes adaptive design features like hints, prompts, and levelling to help students regulate their inquiry processes and encourage efficient and effective information gathering.

Other interesting examples relate to students' fact-checking and information sharing skills in open, networked environments. Games like 'Fake It To Make It' (Urban et al., 2018), 'Bad News' (Roozenbeek and van der Linden, 2019) or 'Go Viral!' (Basol et al., 2021) teach players common techniques for promoting misinformation in the hope that this prepares them to respond to it. In 'The Misinformation Game', participants can engage with posts in ecologically valid ways by choosing an engagement behaviour (with options including liking, disliking, sharing, flagging, and commenting), and are provided with dynamic feedback (i.e. changes to their own simulated follower count and credibility score) depending on how they interact with reliable or unreliable information (van der Linden et al., 2020).

**Source:** Piacentini and Foster (2023), Chapter 3 in *Innovating Assessments*.

### **Contexts of practice or domains of application**

While 21st century competencies are widely seen as being transversal or interdisciplinary, what it means to problem solve, think critically, or be creative in one context may be very different in another context. These skills are neither exercised nor observed in a vacuum, and we can hardly assess them in a domain-neutral way. Hence, when defining the focus of an assessment, the role and importance of domain-specific knowledge should be made explicit from the outset. In an assessment context, students' ability to perform these skills will always be observed in a given context or situation, 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. Attempting to design completely decontextualised problems or scenarios also threatens validity: if a student does not require any knowledge to solve a task, can an assessment truly claim to measure the types of complex problem-solving competencies it claims to be interested in?

Next-generation assessments can be contextualised in a specific domain of knowledge or cross multiple disciplines. Cross-disciplinary here does not mean domain-general, as the competencies that students show on cross-disciplinary tasks still depend on a well-defined set of knowledge; only that knowledge is not limited by the bounds of a single discipline. The most-widely used assessments of learning outcomes are set in one single discipline (e.g. mathematics, biology,

history) and focus on the reproduction of acquired knowledge and procedures relevant to that discipline. When thinking of an assessment of 21st century competencies in the context of a disciplinary domain, new assessments could bring a better balance between the testing of disciplinary knowledge and the evaluation of students' capacity to apply this knowledge in authentic contexts and to new problems. Assessments could invite students to engage in practices that reflect how disciplinary knowledge is used to address both professional and everyday problems. In history, for example, students could be asked to collaboratively investigate and find biases in an historical account of an event. In science, an assessment could ask students to engage in an exploration of a scientific phenomenon in a virtual lab, using relevant tools and progressing through the sequence of decisions that real scientists follow in their professional practice (see Box 2 for a more detailed example).

#### BOX 2.

#### DOMAIN-SPECIFIC ASSESSMENTS OF COMPLEX SKILLS

##### Testing student decision making in the fields of science and engineering

Complex problem solving, particularly in science and engineering fields, is a core competency of the modern world, and many newer science standards have it at their core. However, student assessments do not commonly capture the key processes and decisions that problem solving involves in real life and remain therefore limited to make meaningful conclusions about student competencies.

Solving the types of problems typically found in school exams and textbooks requires recognising and following a single, well-established procedure. These problems can be complicated, in that they require multiple steps, but very few decisions are involved – one either knows the correct procedure or not. This is not how complex problem solving works. Expert scientists and engineers are not experts because they are good at following a specific procedure or technique, but because they are good at applying their knowledge and technical skills to solve problems for which there is no complete information and a defined set of solution steps is lacking. Unlike “school problems”, real-life problems have a mixture of relevant and irrelevant information, and some of the most challenging aspects of solving them relate to addressing questions like What information is needed?, What concepts are relevant?, What is a good plan?, What conclusions are justified by the evidence?.

Wieman and Price (2023) argue that school (and therefore assessment) problems should look more like authentic problems: they should provide students with opportunities to engage and practice the type of decision making that practitioners face in the real world, i.e. learning how to think and reason like a scientist or an engineer. A problem can be authentic, requiring solvers to make decisions instead of following a prescribed procedure, and be constrained to require the knowledge expected of students at a particular level. The key is to have a good understanding of the decisions practitioners face (cognition vertex) and use this knowledge to inform the design of tasks and scoring methods.





Finding the appropriate balance between authenticity and practicality in assessment involves choosing tasks and questions that constrain the problem solver at an appropriate level. Too much constraint means the important resources and decision processes will not be probed, while too little constraint results in responses that can vary so much that it is impossible to evaluate and compare the detailed strengths and weaknesses of the test takers.

**Source:** Wieman and Price (2023), Chapter 4 in *Innovating Assessments*

While incorporating choice and authentic problems within disciplinary assessments represent important avenues for innovating current assessment practices, situating next generation assessments across several domains could also be a valuable approach. One way to engage students in cross-disciplinary tasks could be to propose assessment situations where they have to act as responsible citizens, confronting problems involving a group of peers, a neighbourhood, or wider communities. Modern simulation-based assessments can incorporate many of these experiential learning situations, affording opportunities to make social choices and develop empathetic understanding by projecting oneself through an avatar (Raphael et al., 2009). These contexts may be particularly suited to assess socio-emotional skills such as communication, cooperation, emotion regulation and empathy. An increasing number of role-play games have been designed to assess these skills in a stealth way, such as *Hall of Heroes* (Irava et al. 2019). All the same, a significant challenge in developing cross-disciplinary assessments is that we lack solid theories about the development of knowledge and skills in these “domains”. Defining exactly what factors are construct-relevant or irrelevant, and what constitutes ‘good performance’ in a cross-culturally valid way are related challenges.

### **Individual vs collaborative tasks**

Group work is increasingly used as a pedagogical practice across the world, despite the challenges for teachers to effectively structure and moderate collaborative learning (Gillies, 2016). Researchers and teachers have become increasingly aware of the positive effects that collaboration might have on students’ ability to learn. Research shows that collaborative work promotes both academic achievement and socialisation abilities, and these positive effects hold across age and disciplines (Baines et al., 2007; Gillies and Boyle, 2010). Formative assessment practices have followed this trend, as more teachers around the world apply rubrics to evaluate their students’ capacity to work in groups. In summative assessments, progress has been much more hesitant, although with notable exceptions (see Box 3 for two examples in large-scale assessments).

BOX 3.

### ASSESSING STUDENT COLLABORATION IN LARGE-SCALE ASSESSMENTS

#### The cases of PISA 2015 and the Assessment and Teaching of 21st Century Skills (ATC21S)

Within the framework of the PISA assessment of collaborative problem solving, three competencies form the core of the collaboration dimension: establishing and maintaining shared understanding, taking appropriate action to solve the problem, and establishing and maintaining group organization. The ATC21S identifies similar dimensions of collaboration: participation, perspective-taking, and social regulation.

There is a key difference between these two experiences: in PISA, students interacted with computer agents, while ATC21S opted for human-to-human collaboration. PISA's choice was justified by the goal of standardising the assessment experience to enable the use of established scoring methods. The interaction between students and the agent was limited to pre-defined statements using a multiple-choice format, and every possible intervention of students was attached to a specific response by the computer agents or event in the problem scenario. This highly controlled test environment and the lack of open response formats for students inevitably reduced the authenticity of the assessment.

In contrast, the human-to-human approach of ATC21S has more face validity, as students could choose when and how to interact with peers using a chatbot. However, in this more open environment, the behaviour of students is difficult to anticipate, and this creates obvious challenges for scoring. Additionally, the success of one student depends on the behaviour of other students, as well as the stimuli and reactions that they offer: this generates the measurement problem of how to build separate scores for students and for their group, and raises the concern of whether it is fair to penalise a student for the lack of ability or motivation of another student.

These experiences suggest that it is possible to imagine a not-so-distant future in which collaborative tasks are an integral component of assessments. Hu, Shubeck and Sabatini (2023, Chapter 10 in *Innovating Assessments*) provide examples of how natural language processing (NLP) can be leveraged to increase the authenticity of the interaction with the virtual agents by designing intelligent agents that 'understand' what students write or say and respond accordingly. Similarly, advances in NLP have the potential to enable the automated replication of expert judgements to large sets of conversational data, improving the quality and reducing the costs of analysis of recorded conversations and written chats between student peers. Regardless of the approach, the realisation of authentic collaborative tasks requires substantial parallel innovation in measurement, as standard analytical models cannot deal with the many interdependencies across time and agents that arise in collaborative settings.

**Source:** Piacentini and Foster (2023) and Hu, Shubeck and Sabatini (2023), Chapters 3 and 10 in *Innovating Assessments*.

## ESTABLISHING SOLID CONCEPTUAL FOUNDATIONS

With greater clarity on the target activities, contexts, and agents for a new assessment, it is then necessary to make an inventory of the concepts, language, and tools that people use in the target domain and define the characteristics of good performance in those domain contexts. In traditional assessments of disciplinary subjects (e.g. maths), detailed descriptions of the domain are already available for use in assessment design. For instance, if we want to assess reading ability, assessment developers can rely on an extensive literature that defines the knowledge and skills required and that has examined how children learn to read and progress in proficiency. However, the same understanding or knowledge on learning progressions is not available for complex competencies like collaborative problem solving or communication.

To generate such information, assessment designers may rely on the contribution of a group of experts who are capable of constructing new representations of what expertise means in those domains, using empirical observations as much as possible. Cognitive task analysis (CTA) uses a variety of interview and observation strategies, including process tracing, to capture and describe how experts perform complex tasks (Clark et al. 2008). For example, an established strategy used for CTA is the critical incident technique, in which an expert is asked to recall and describe the decisions they made during an authentic situation (see Chapter 4 in *Innovating Assessments* for an example of this practice). The descriptions generated through CTA are then used to develop training experiences and assessments, as they make it possible to identify features of tasks that are appropriate to include and identify decisions that are most indicative of competence.

Defining an empirically-based model of the domain can be supported by observational studies of how students work on tasks that engage the target skills. For example, in an assessment of collaboration skills, developers can craft some model collaborative activities that reflect their initial understanding of relevant situations in the domain. They can then use CTA methods to identify those students who are more or less successful in driving the collaboration towards the expected outcome and make an inventory of what students at different proficiency levels say and do (e.g. how they share information within a group, how they negotiate the sharing of tasks, etc.). Observational studies provide clarity on the sequence of actions that must be performed in order to achieve a performance goal and produce exemplars of real work products or other tangible performance-based evidence that can be associated to proficiency claims.

In the subsequent phases of development, assessment designers collaborate with domain experts to organise the information collected in their domain analysis into assessment arguments, i.e. the claims they want to make on student performance, the data that will serve as evidence for such claims, and the warrants, or reasons, that explain why certain data should be considered appropriate evidence for a certain claim (Toulmin, 1958; Mislevy and Riconscente, 2006). As exemplified in Box 4, assessment arguments can be usefully formalised using “design patterns”, which describe the student knowledge, skills and attitudes that are the focus of the assessment, the potential observations, work products and rubrics that



test designers may want to use, and the characteristics of potential assessment tasks. This design pattern structure helps to identify consolidate the conceptual foundations of an assessment and serves as the basis to elaborate the technical specifications guiding the operationalisation of the assessment – that is, the student, task and evidence models of the ECD framework in Figure 3.

BOX 4.

#### DESIGN PATTERNS IN ASSESSMENT: AN EXAMPLE FROM PISA

##### Design pattern for computational modelling in the PISA 2025 Learning in the Digital World assessment

<b>Rationale (warrant)</b>	Modelling is a core practice in scientific reasoning, but students rarely engage in modelling during compulsory education. Computers make modelling more accessible and meaningful to learners, in particular novices. Observing how students build, refine, and use computational models provides relevant and interpretable evidence on how capable students are to create their own knowledge and understanding of complex phenomena using computers.
<b>Focal knowledge, skills, and attitudes</b>	<ul style="list-style-type: none"> <li>Understanding the concept of variables, including dependent, independent, control and moderating variables</li> <li>Creating an abstract representation of a system that can be executed by a computer; ensuring that the model functions as expected (e.g. observing behaviours of agents in a simulation based on the model)</li> <li>Identifying trends, anomalies, or correlations in data</li> <li>Experimenting using the control-of-variables strategy</li> <li>Using a computational model to make predictions about the behaviour of a system</li> </ul>
<b>Additional knowledge, skills, and attitudes</b>	<ul style="list-style-type: none"> <li>Functional knowledge of ICTs</li> <li>ICT self-efficacy</li> <li>Prior knowledge of the phenomenon to be modelled</li> <li>Perseverance, conscientiousness, and mastery orientation</li> </ul>
<b>Potential observations and work products</b>	<ul style="list-style-type: none"> <li>Student model represents the available information on the real-world situation</li> <li>Student consults relevant information resources and collects relevant data to set the model parameters</li> <li>Student modifies an incomplete or faulty model, and justifies their modifications</li> <li>Student identifies model weaknesses</li> <li>Student uses their model to make correct predictions (given the available data)</li> </ul>
<b>Characteristic features of tasks</b>	<ul style="list-style-type: none"> <li>Students are either provided with information about a real social or scientific phenomenon to model or provided with the tools to obtain this information.</li> <li>The student can check their model by comparing its output with real data</li> <li>Students can use the model to make predictions</li> </ul>
<b>Variable features of tasks</b>	<ul style="list-style-type: none"> <li>Level of familiarity of the phenomenon to model</li> <li>Complexity of the ICT tools used for modelling</li> <li>Student improves a basic model (provided to them) or builds the model from scratch</li> <li>Student must find relevant data (e.g. in an information resource) or generate their own data through experimentation</li> <li>Number of variables to be modelled and structure of the system (simple vs. multi-level)</li> </ul>
<b>Constraints and challenges</b>	<ul style="list-style-type: none"> <li>Limited time to learn how to use the modelling tool</li> <li>Limited time to learn unfamiliar modelling concepts (e.g. control of variable strategy)</li> <li>Large differences in prior knowledge in the target student population, meaning difficult to appropriately challenge all students on the same task</li> </ul>

**Source:** Piacentini (2023), Chapter 6 in *Innovating Assessments*.

## **CONSIDERING SOCIOCULTURAL DIFFERENCES WHEN DEFINING ASSESSMENT CONSTRUCTS**

Comparative inferences require equivalence of measurement and comparability of scores when tests are administered in multiple languages or when students from different cultural groups take tests in the same language. Cross-cultural validity and comparability issues have particular relevance to assessments of complex constructs in multicultural and multilingual contexts, such as in international assessments, and assessments in countries with culturally diverse populations.

Construct equivalence is an important aspect to pay attention to in defining the focus of an assessment. The equivalence of constructs is the degree to which definitions of constructs are similar for populations targeted by the assessment, whether individuals are expected to develop and progress on these constructs in similar ways, and whether the constructs are accessible in similar ways for all populations. It is critical to all assessments intended for multicultural and multilingual groups, but it takes on specific relevance to large-scale assessments of complex and multidimensional constructs (Ercikan and Oliveri, 2016).

Constructs such as creativity, intelligence, critical thinking, and collaboration are not uniformly taught in schools, and are conceptualised and defined differently in different cultures. For example, how creativity develops and how creative behaviours are manifested differ across cultural groups (Lubart, 1990; Niu and Sternberg, 2001). Other researchers have also argued that the concepts of intelligence are grounded in cultural contexts and, as such, the constructs have different definitions in these contexts (Sternberg, 2013).

Given that complex skills are embedded within social contexts and are characteristically shaped by cultural norms and expectations, we can expect their manifestations and the value attributed to student outputs to vary across cultures. Because of these differences across cultural groups, there is a need to clearly evaluate what aspects of a construct can be meaningfully assessed in a comparative context and thus included in the assessment, even if this might result in some narrowing of the construct. The PISA 2022 assessment of creative thinking (OECD, 2022) exemplifies how an assessment of a complex construct across language and cultural groups can nonetheless focus on certain aspects of the construct that optimise comparability (see Box 5).

BOX 5.

### **CONSIDERING SOCIOCULTURAL DIFFERENCES IN CONSTRUCT DEFINITION OF LARGE-SCALE ASSESSMENTS**

#### **Ensuring construct equivalence in the PISA 2022 Creative Thinking Assessment**

The PISA 2022 Creative Thinking Assessment emphasises that assessment items should draw upon knowledge and experiences that are common to most students around the world and for which students can meaningfully and realistically produce creative work within the constraints of a PISA environment.

To ensure this, the assessment developers considered five issues in particular:

- Focused the assessment on the narrower construct of creative thinking (rather than on the broader construct of creativity), defined as the “competence to engage productively in the generation, evaluation and improvement of ideas”. This narrower focus emphasised the cognitive processes related to idea generation, whereas creativity also encompasses personality traits and requires subjective judgements about the creative value of students’ responses.
- Defined creative thinking, how it is enabled (i.e. indicators of opportunities to learn creative thinking), and what it looks like in the context of 15-year-olds in the classroom, focusing on aspects of the construct that would be more likely to be developed in schooling contexts (rather than outside of school).
- Identified cross-culturally relevant assessment domains in which 15-year-olds could engage and could be expected to have practiced creative thinking (e.g. writing short stories, creating visual products, brainstorming ideas on common social and scientific problems).
- Focused on the originality of ideas (defined as statistical infrequency) and on the diversity of ideas (defined as belonging to different categories of ideas) in scoring, rather than their creative value (considered more likely to be subject to sociocultural differences).
- Engaged in significant cross-cultural verification of the coding rubrics that human raters used to evaluate the responses, including refining those rubrics through the analysis of sample responses from students in several countries.

**Source:** OECD (2022), Thinking Outside the Box: The PISA 2022 Creative Thinking Assessment, <https://www.oecd.org/pisa/innovation/creative-thinking/>.

## **INNOVATING THE OBSERVATION VERTEX: INCLUDING MORE VARIED AND INTERACTIVE ASSESSMENT TASKS**

From the perspective of assessment as a process of reasoning from evidence, assessment tasks must elicit relevant evidence from students, and this evidence needs to be clearly connected to the construct. In other words, assessment tasks or situations should allow for the observation of the types of performances we expect students to master. For constructs like mathematics knowledge, the link between test indicators and construct is fairly direct: a correct response to a given question demonstrates knowledge of the topic. But this logic may not be sufficient to capture the complexity of 21st century competencies.

A central argument of *Innovating Assessments* is that assessments are more likely to generate valid evidence of what students know and can do if they confront students with authentic situations. Motivating the call for innovation is the fact that existing assessments do not often make this possible, in part because the technical capabilities to instantiate such a vision at scale have been slow to emerge. Educational assessments, particularly large-scale standardised tests, have been designed within a set of constraints – printing and transporting costs, test security, test environment, testing time, and cost of scoring – while needing to satisfy psychometric standards of reliability, validity, comparability and fairness. The main features of “traditional” test design, administration, scoring, and reporting, such as multiple-choice items, have taken shape because of such constraints (OECD, 2013), and their capacity to capture more complex and multi-faceted aspects of performance has remained consequently limited.

All the same, many of the constraints in test design and administration either no longer apply, have been transformed, or can be relaxed in large part due to technological and data analytic advances. In particular, the digital toolbox available to test developers now dramatically expands assessment design opportunities and affordances, with the potential to make test experiences less artificial and more face valid by approximating or simulating the situations or contexts in which target constructs are used in real life.

## RETHINKING TASK DESIGN

Piacentini, Foster and Nunes (2023) provide a set of design innovations for tasks and items (Chapter 2 in *Innovating Assessments*). These include (1) allowing for extended, performance tasks with “low floors” and “high ceilings”; (2) explicitly accounting for domain knowledge; and (3) providing opportunities for productive failure and learning on the test, offering feedback and instructional support during the assessment. The idea behind these principles is not to get rid of more traditional forms of assessment experiences and response formats, as those can still provide relevant information for some interpretative uses (e.g. identifying knowledge gaps). Rather, the argument is to complement those established forms of assessment with a different set of assessment experiences that incorporate these innovative features.

### **Design principle #1: Allow for extended, performance tasks with “low floors” and “high ceilings”**

In assessment, particularly large-scale summative assessments, efficiency considerations have led to the primacy of short, discrete assessment tasks over longer performance activities. In general, using many short items provides more reliable data on whether students master specific knowledge and can execute a set of given procedures, as the information is accumulated over a larger number of observations. Measurement is also easier: the evidence is accumulated by applying established psychometric models to items that are fully independent. However, if the purpose of assessment shifts to evaluating whether students can construct new knowledge in choice-rich environments, then students should be presented with assessment tasks and environments that are appropriate for this goal.

To do this, it is important to consider how an assessment can provide students with a challenge that is purposeful and that allows enough time for them to demonstrate their competencies. Including extended units, where multiple activities are sequenced as steps towards achieving a main goal, can provide students with a more authentic and motivating assessment experience. Encouraging a shift in the test taker’s mindset, from ‘I have to get as many of these test items right’, to ‘I have a challenge to work towards and accomplish’, might ultimately provide more valid evidence of what students are capable of doing outside of the constraints of stressful and time-sensitive test contexts.

Extended, performance-based tasks are more challenging to design, not least because one needs to establish a coherent storyline that keeps students engaged and to address potential dependency problems – for example, by providing rescue points to move struggling students from one activity to the next. At the same time, assessments should allow all students to demonstrate their ability to learn and progress, regardless of their initial level of knowledge or skill by designing tasks that have ‘low floors’ and ‘high ceilings’, meaning that they are accessible to all students while still challenging top performers (see Box 6 for an example from OECD’s PISA platform).

One way to engineer low floor, high ceiling problems is to ask students to produce an original artefact: this could be a story, a game, a design

for a new product, an investigation report on some news, a speech, etc. These more open performance tasks generate a wide range of qualitatively distinct responses, and even top performers have incentives to use resources that can help them produce a solution that is richer, more complete and unique. The low floor, high ceiling design can also be used in the context of more standardised problem-solving tasks, making clear to students that there are intermediate targets to achieve and that they are expected to progress as much as they can towards a sophisticated solution.

Adaptive designs can also address the complexity of measuring learning in action amongst heterogenous populations of students. A relatively simple way to do it involves creating scenarios where students have a complex goal to achieve, and they progress towards this goal by completing a sequence of tasks that gradually increase in difficulty (similar to 'levelling-up' in videogames). More proficient students will quickly complete the initial set of simple tasks, after which they will encounter problems that challenge them; and less prepared students will still be able to engage with the simpler tasks, even if they do not complete the full sequence. Within such designs, both groups of students work at the cutting edge of their abilities, with obvious benefits in terms of measurement quality and test engagement. With current technologies, this design could be further improved by introducing multiple, adaptive paths within a scenario: based on the quality of their work, students are directed on-the-fly towards easier or more difficult sub-tasks.

#### BOX 6.

#### **ASSESSMENT TASKS WITH 'LOW FLOORS' AND 'HIGH CEILINGS'**

##### **Catering to differently able students in the PILA assessment of computational problem solving**

The Platform for Innovative Learning Assessments (PILA) is a research laboratory coordinated by the OECD. The assessments in PILA are designed as learning experiences and provide real-time feedback on students' progress. They can thus also be used in the context of classroom instruction. An 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 application developed in PILA focuses on computational problem solving. Students use a block-based visual programming interface to instruct a turtle robot ("Karel") to perform certain actions. The assessment has a low floor and high ceiling: the intuitiveness of the visual language and the embedded instructional tools (e.g. interactive tutorial, worked examples) allow students who have no programming experience to engage successfully with simple algorithmic tasks. However, the same environment can also be used to create problems that can challenge even expert programmers.

The images below show an example problem asking students to build a single program that lets Karel achieve the goal in two different



scenarios. To solve the problem, students are able to toggle between the two scenarios to visually observe the differences in the environment, as well as the extent to which their program solves the problem in both scenarios. In these kinds of tasks, even students with a solid programming background generally develop and run multiple iterations of their program before finding a solution. The scoring models takes into account partial solutions (e.g. a student's ability to solve the problem in one world), and the reporting dashboards include more complex indicators of performance (e.g. the number of iterations students tested).

**Challenge:** Fill in the missing code in the main function to help Karel achieve the goal for both Scenarios.

**Start:**  **Goal:** 

Scenario 1: Not Tried     Scenario 2: Not Tried

Play Speed: (slow) — (fast)

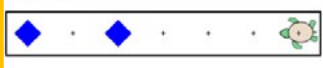

**Code:**

```

define main
  while front is clear
    move forward
    if front is clear
      place stone
    pickup stone
  if front is clear
  while front is clear
  
```

**Task:** the students need to program Karel to move forward and place one stone along the way so to match scenario 1's goal state (image above). The same code should also solve scenario 2 (image below).

**Challenge:** Fill in the missing code in the main function to help Karel achieve the goal for both Scenarios.

**Start:**  **Goal:** 

Scenario 1: Not Tried     Scenario 2: Not Tried

Play Speed: (slow) — (fast)

**Code:**

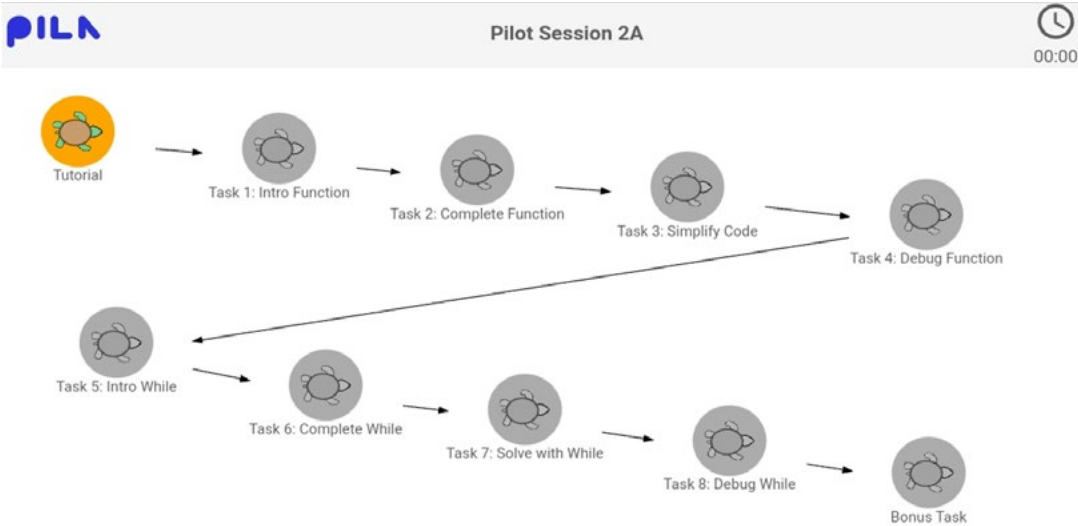
```

move forward
turn left
place stone
pickup stone
if front is clear
while front is clear
  
```



Each PILA assessment experience is also structured as a progression of increasingly complex tasks that have a common learning target (e.g. using functions efficiently). Assessment designers and teachers have the option of locking students in a particular task until they are able to solve it (i.e. a 'level-up' mechanism) or students are able to control how they move along the task sequence. Only highly skilled students are expected to finish the whole task sequence, and this is communicated clearly to students at the beginning to reduce experiences of frustration. In the future, PILA plans to include adaptive pathways (i.e. problem sequences that adapt in real-time to student performance), in order to further align the experience with the students' previous knowledge and skills.

Example of Assessment



Experience ('Map') in the Karel application.

**Source:** Piacentini, Foster and Nunes (2023), Chapter 2 in Innovating Assessments



### **Design principle #2: Explicitly account for domain knowledge**

As previously discussed, when designing assessments of 21st century competencies, it is important to explicitly identify the knowledge students need 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 skills. In the context of large-scale, summative assessments, it might be misleading to make general claims such as ‘students in country A are better problem solvers than students in country B’. In fact, from a single summative assessment we might only claim that students in country A are better than students in country B at solving problems in the situations that are presented in the test (most likely, a limited number of situations contextualised in one or few domains of knowledge).

Measuring the relevant knowledge that students have when they undertake a performance task (for example, through a short battery of items at the beginning of the test) should become an integral part of the design and assessment process in next-generation assessments. This information can also help to interpret student’s behaviours and choices in assessments with complex performance tasks. Assessments could also seek to minimise variability in relevant prior knowledge by providing students with tutorials, examples, and walkthrough problems that can help them engage with a task. These approaches can be useful both for accounting for domain knowledge, but also knowledge about the resources or tools embedded in the assessment environment (i.e. helping students navigate the test environment).

### **Design principle #3: Provide opportunities for productive failure and learning on the test, offering feedback and support mechanisms**

In traditional tests, the goal is to assess students’ acquired knowledge prior to the task. Usually no feedback is given to students, tasks are likely very distinct from one another (to avoid giving away the answers in the same test), and the types of responses are mostly limited to categorical responses, i.e. correct or incorrect answers. These instruments are insufficient when assessment goals expand from evaluating the application of existing, static knowledge (learning *outcomes*) to evaluating the dynamics of acquiring and developing new knowledge (learning *processes*) when facing complex tasks.

One promising method to address current shortcomings involves the use of ‘invention activities’ in assessment, which ask students to solve problems that are seemingly unrelated to the class material and that involve concepts or procedures that they have not yet been taught. Students have to invent their own original solutions to these novel problems, and in this process, they tend to make mistakes and fail to generate canonical solutions. However, invention activities help students to deeply understand the concepts, let go of old interpretations and procedures when they do not work, and look for new patterns and interpretations – and in the context of an assessment, can provide evidence on whether students can flexibly apply their knowledge schema to unfamiliar contexts as adaptive experts do. Certainly, fully open and unguided exploration may not provide the most useful evidence of what novices can do; learning activities must nonetheless be carefully designed to support students in building their understanding as they invent and interact with problems that have unfamiliar aspects.

Next-generation assessments should consider including guidance and scaffolding during the solution process in the form of advice, feedback or prompts. Such scaffolding can play a variety of functions: (1) engaging student's interest when they appear disengaged; (2) increasing their understanding of the requirements of the task when they demonstrate confusion; (3) reducing degrees of freedom, or the number of constituent acts required to reach a solution; (4) maintaining direction; (5) marking critical features, including discrepancies between what the student has produced and what they would recognise as correct; (6) demonstrating or modelling solutions, for example reproducing and completing a partial solution attempted by the student; and (7) eliciting articulation and reflection (Guzdial, Rick and Kehoe, 2001).

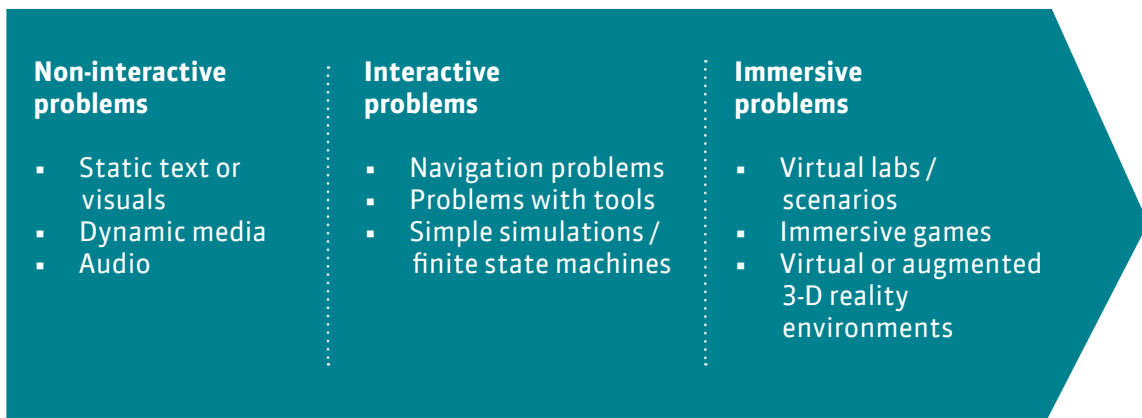
### LEVERAGING MODERN TECHNOLOGIES TO INNOVATE ASSESSMENT DESIGN

Ongoing technological developments make the abovementioned design innovations increasingly feasible by expanding the tools available for assessment design. As discussed by Sabatini and colleagues (*Innovating Assessments*, Chapter 7), modern technologies expand the range of the possible when it comes to designing task formats, test features, and sources of evidence.

#### Task format: From static to interactive and dynamic assessment situations

Many assessments are characterised by non-interactive problems. These often include static written text or visual stimuli (e.g. photos, drawings, tables, maps, graphs, or charts), and in some cases, more dynamic stimuli like audio, animations, video, and other multi-media content. In non-interactive problems, stimulus material usually provides students with all the information they need to solve the task, responses often take the form of written or close-ended items with little to no test taker interactivity possible, and the test environment does not evolve as the test taker interacts with it.

**Figure 5.** Task format continuum  
Non-interactive, interactive and immersive assessment problems



**Source:** Sabatini et al. (2023), Chapter 7 in *Innovating Assessments*.

In contrast, interactive problems allow students to engage actively in the processes of making and doing by creating problem-solving scenarios that characterise more complex types of performances. These types of task format are more open and responsive to test takers actions and behaviours. They are typically multi-step, involve the use of computer applications, tools or search engines that better reflect contemporary contexts of practice, and usually require navigation within and across screen displays.

Assisted by technology, assessments can also incorporate truly immersive problems. These include simulated labs, immersive games, or 3-D modelling and virtual reality environments. Immersive problems allow examinees to navigate through a two- or three-dimensional rendition of a virtual world – imaginary or real – on a screen or via virtual reality headsets. Immersive problems frequently employ game-based elements to enhance motivation, as well as scaffold or control learner experience (Pellas et al., 2018). Examples include simulations used most often for professional training, such as virtual aviation or medical intervention simulations, although these types of tasks are increasingly becoming feasible to design and implement at scale.

Importantly, greater interactivity and immersivity in assessment tasks needs to be balanced with construct and practical considerations. Tasks that evolve as test takers interact with them may result in less uniform task experiences and therefore uneven coverage of the target constructs, creating challenges in drawing inferences across student populations. Authentic and interactive tasks might also take more time to complete than simpler, static tasks. Task design for interactive and immersive problems necessarily involves optimising the trade-off between task authenticity and constraints: in immersive designs, it is paramount that tasks in the virtual world are sensitive to variations in performance between individuals (e.g. real-world novices and experts), that they truly reflect the knowledge and skills of interest (i.e. that they have construct validity), and that they do not adversely distract students from the task at hand.

### **Test features: Introducing test adaptivity and learning resources**

Digital technology can also serve to innovate test features, which refer to the affordances or characteristics that can be overlaid with any of the abovementioned task formats. Two types of features are particularly considered here for next-generation assessments: adaptivity, and learning resources.

First, digital test delivery has enabled computer adaptive testing (CAT). One of the most researched innovations in test design (e.g. Wainer et al., 2000), decision rules or algorithms select test items from an item pool for individual examinees, and while different examinees may take different items or larger modules in the same assessment, their scores are placed on a common scale and remain comparable. In general, test adaptivity increases efficiency, accuracy and fairness in assessment design, administration and interpretation, although different CAT designs have different strengths and weaknesses (see Box 7).

BOX 7.

## COMPUTER ADAPTIVE TESTING: POSSIBILITIES AND CHALLENGES

### Strengths and weaknesses of different CAT designs

Various CAT designs have been researched and implemented in large-scale assessment. In simpler adaptive designs, test items are grouped into modules that differ in difficulty and a computer algorithm directs students to one module or another depending on performance. Tests may include multiple stages, and stages include several modules (depending on module and test length). Different algorithms can be used to make branching decisions between stages. In these designs, adaptivity occurs at the stage-level. Other designs employ on-the-fly adaptivity, where adaptivity occurs at the item-level (i.e. each item is tailored to the student based on their performance on previous items). One advantage of single- or multi-stage adaptive testing (MSAT) over item-level adaptive testing is that it allows modules to include larger and more complex task formats that have their own internal naturalistic logic for items contained within the task. Conversely, on-the-fly approaches where test forms are not defined a priori by test developers but defined during the testing time by the computer are efficient in the delivery of items for the given constraint set, and may provide a more precise estimate of ability per unit of test time. The weakness of basing next item decisions solely on performance is that it can result in reduced construct coverage and in an arbitrary (rather than cohesive or thematic) trajectory through the content domain. Recent advances in CAT may help to address this issue by integrating hybrid measurement models, although these designs are far less mature than their well-researched counterparts.

A different CAT design adapts tasks based on prior choices or actions of the examinee – like videogames do based on players' actions and behaviours. This approach has the advantage of better reflecting the contingencies in real problem-solving environments and, if designed to give examinee some choice or control, can enhance engagement. However, enabling adaptivity fully based on test taker choice can introduce construct-irrelevant variance when choice is not explicitly part of the assessment framework. Even in cases where choice is explicitly assessed, similar issues can arise as with on-the-fly adaptive models without sufficient constraint mechanisms. Internally adaptive tasks also require complex algorithms to deliver. Techniques for developing such designs quickly and efficiently have not yet emerged, rendering this type of adaptivity expensive to develop and pilot and more difficult to score for the purpose of standardised assessment. However, innovative assessments might be able to integrate this type of more complex, multi-level adaptivity by adopting some of the technical solutions already used in videogames that are designed to maintain player engagement, for example alternating states of learning and states of mastery.

**Source:** Sabatini et al. (2023), Chapter 7 in *Innovating Assessments*.

Second, digital technologies facilitate the inclusion of learning resources in assessments. When the focus of assessment is just on measuring how well students know or can do something at a given point in time, then there is no need to incorporate learning resources in the task. However, innovative assessments might want to make claims about how students deal with authentic problem situations, how they adapt their problem-solving strategies as they increase their understanding of a problem, and how they do so by using a diversity of resources. Box 8 describes three types of affordances that become possible with the integration of learning resources in technology-rich assessments.

BOX 8.

### **INNOVATING TASK DESIGN WITH LEARNING RESOURCES**

#### **Three types of affordances in technology-rich assessments**

Learning resources offer multiple affordances to enact goal-oriented behaviours. Roll and Barhak-Rabinowitz group such affordances into three families: Experimentation, Explicit-feedback, and Information-seeking.

Experimentation allows learners to interrogate and represent their ideas and execute them in a manner that produces responses from the environment. For example, coding environments let students code, compile, execute, and observe the outcomes (conversely, coding tasks where learners enter code but cannot execute it are not considered learning resources according to this definition). Another example is interactive scientific simulations where learners can manipulate elements and observe the outcome of their exploration (e.g. Wieman, Adams and Perkins (2008)). The main benefit of experimentation resources comes from their responses to learner actions, often termed situational feedback (Nathan (1998); Roll et al. (2014)). For example, an interactive simulation for electricity will adjust the shown light intensity based on the voltage that learners set (de Jong et al. (2018); Roll et al. (2018)). Situational feedback is implicit and originates within the task situation itself, consistent with the internal logic of the task. That is, learners are not being flagged or graded by an external all-knowing model. Instead, they are given opportunities to elicit, observe, and interpret the relevant information from the environment response (Nathan, 1998). Observing how learners respond to situational feedback can be used to evaluate their monitoring behaviours and the corresponding adjustments that they make in their cognitive strategies.

Explicit Feedback affordances provide learners with an evaluation of their actions. This can include a range of inputs, from error flagging to explanations about the nature of error or suggestions for future work (Deeva et al., 2021). Feedback can be triggered on-demand (e.g. using a “test” button) or automatically (e.g. following a set number of failed attempts). Unlike situational feedback that is built into the narrative of the challenge, explicit feedback is external. It assumes an “all-knowing” agent or environment that can compare the student input to the desired state. The use of on-demand explicit feedback offers a direct measure of learners’ metacognitive strategies such as monitoring, or which sub-goals they pursue (Winstone et al., 2016). As with situational feedback, students who choose to adjust their cognitive strategies



effectively following explicit feedback demonstrate productive use of metacognitive strategies (e.g. Kinnebrew, Segedy and Biswas (2017)). Information-seeking affordances support learners by providing additional communication about the task at hand. Informational resources include hints (e.g. Aleven et al. (2016)), instructional videos (e.g. Seo et al. (2021)), worked examples (Ganaiem and Roll (2022); Glogger-Frey et al. (2015)) searchable databases, etc. Information sources can be fixed (as in most tutorials) or adaptive (as in hints about the specific problem step; VanLehn et al. (2007)). When using information sources, learners make choices regarding when to use them (e.g. when to ask for hints), how to use them (e.g. navigating videos), and how to apply the information to the challenge at hand. Effective and strategic learners seek just-in-time information to fill their own knowledge gaps (Seo et al. (2021); Wood (2001)). Thus, interactions with information resources can provide meaningful insights into learners' help-seeking and monitoring processes (Roll et al., 2014).

For any of the above, it must be noted that enabling choice in the context of providing learning supports integrates an additional construct in the assessment. Choice therefore needs to be explicitly reflected in the definition of the domain and incorporated into inferences about examinee performance.

**Source:** Roll and Barhak-Rabinowitz (2023), Chapter 9 in *Innovating Assessments*

Decisions over the exact type and nature of support provided to students should be guided by the goals of the assessment as specified in the assessment framework (the cognition vertex). For example, where the use of feedback is considered construct-relevant, intelligent feedback mechanisms embedded into tasks should always be useful to students. In other words, if all students receive the same feedback but this is not useful to some of them, these may not be able to demonstrate the targeted skill. Similarly, where test taker choice is construct-relevant, perhaps an on-demand mechanism is appropriate; however, enabling choice may also preclude opportunities to observe such behaviours, so it may be desirable to build in some action- or event-triggered feedback mechanisms as well.

A key challenge of introducing learning resources in assessment relates to deciding which scoring models to apply. These support systems can potentially change the knowledge state of the examinee as the test proceeds, influencing examinees' performance on future test items. What remains is for the extensive research that has been conducted on feedback, scaffolding and resources as learning devices to be conducted in psychometric modelling for assessment design. For example, where examinees are given more than one chance to respond (e.g. after receiving feedback), scoring models might weight answering correctly the first time higher than subsequent attempts. Alternatively, it may be that reaching a correct answer,

even with supports, warrants full credit. Close interaction with a psychometrics team during the assessment development process is critical for understanding the types of inferences that can be made and how when to integrate such features in the statistical model.

### NEW SOURCES OF EVIDENCE: RESPONSE PRODUCT AND PROCESS DATA

Computer-based tests expand the range of potential evidence sources in assessments. The palette of potential evidence goes well beyond the traditional multiple-choice or constructed (written) responses that have dominated traditional assessment designs, particularly large-scale tests. A key conceptual distinction in this sense is between response products and response processes, and the different types of evidence that these different sources of data generate (see Table 1).

TABLE 1.  
SOURCES OF EVIDENCE

#### Response product data and response process data

PRODUCT DATA	PROCESS DATA
Various selected response (e.g. multiple choice, true/false, drag-and-drop, hotspot, etc.)	Timing data (e.g. time on task, time to first action, inactive time)
Written response	Intermediate solution states (i.e. those before submitting final solution)
Spoken response	Action logs (e.g. use of affordances, keystroke strokes, mouse clicks, events)
Performance response (e.g. level attainment in a game, simulation state, artefact)	Physiological measures (e.g. eye-tracking data)

**Source:** Sabatini et al. (2023), Chapter 7 in *Innovating Assessments*.

Response products refer to students' final responses on an assessment task or a given item; response product data therefore typically refer to data resulting from selected responses (e.g. on a multiple-choice item), short or extended written responses, or the final product in a simulated or real performance demonstration. Response processes rather refer to the thought processes, strategies and approaches of examinees when they read, interpret and formulate solutions to assessment tasks (Ercikan and Pellegrino, 2017). Response processes go beyond the cognitive realm, including emotions, motivations and behaviours (Hubley and Zumbo, 2017). Data that captures potential evidence of these processes can therefore be understood as (response) process data, which typically includes data representing actions or sequences of actions, eye-tracking

data and timing data, as well as data beyond the specific response format, such as in-task chats and dialogues with virtual agents or human collaborators.

The simplest form of product data is generated through selected response formats, like multiple-choice or true/false items presenting pre-defined answers to students. These response formats are easier and cheaper to score than other formats, but test takers may be able to guess the correct answer and, more generally, these formats that cannot provide direct evidence of production skills.

Other forms of product data (constructed responses) can provide this evidence, such as written responses (ranging from short, discrete sentences to extended essays), spoken responses, or via the construction of an artefact or representation (e.g. engaging in a realistic building design in an architecture exam or performing an operation in a medical simulator). By requiring students to engage in a production activity, constructed responses are less susceptible to unduly rewarding students for guessing behaviours and are more suitable for generating evidence of successful learning and problem solving. However, they also require a greater investment on the part of test takers and the data they generate can be more complex to score in a reliable and comparable manner. For example, typical scoring models may take the form of rubrics or guidelines, but these may nonetheless restrict the design of authentic tasks by requiring the kinds of responses for which trained scorers can obtain reliable judgments of quality.

Advances in technology and data analytics (e.g. natural language processing, speech recognition software) are converging to remove some of these barriers. For example, syntactic analytical tools can be used to evaluate the structure of student answers, and machine learning algorithms can be trained to identify semantic similarity between student responses and the answer keys (see Hu, Shubeck and Sabatini (2023), chapter 10 in *Innovating Assessments*).

### **The emergence of response process data**

Besides response products, an outstanding breakthrough in technology-supported assessments is the capacity to generate evidence from response processes. Students' interactions with digital assessment environments can be logged to provide data on how they engage in particular processes, which can be critical to understand the operations that students perform when solving a task and why. Response process data offer the opportunity to reveal these actions, including where and how students spend their time and what choices they make in interactive and immersive environments, which might be useful for making inferences about student thinking (Ercikan and Pellegrino, 2017).

Process data can be exceptionally varied (e.g. online behaviour, gesture and facial expression, verbal interaction, eye movement) and each source of such data can contribute to our understanding of some aspect of how test takers engage with assessment tasks. In this sense, process data can constitute evidence of performance if suitable interpretation methods are employed to make valid infer-



ences, but it can also constitute a highly valuable tool in assessment validation efforts by supporting assessment developers to understand how different students engage with a given assessment environment (see Ercikan, Guo and Por (2023); chapter 12 in *Innovating Assessments*).

## **INNOVATING THE INTERPRETATION VERTEX: MAKING SENSE OF ASSESSMENT OBSERVATIONS**

Previous sections highlighted a growing consensus on the need to focus assessment on what matters; that in order to do measure these more complex competencies, assessments need to provide students with open and interactive assessment problems situated within authentic contexts; and that technology-supported assessments can expand the types of evidence we can rely upon to make measurement claims, including data sources that can elucidate how students think, act, and learn if we have suitable interpretation tools – that is, if we have robust warrants. It is here that lies the third argument calling for innovating assessment: while defining assessment constructs of complex competencies and capturing new forms of evidence is relatively “easy”, with the support of domain experts and digital technology, making defensible interpretations of what the evidence means is far more complicated.

The challenge stems from the fact that the interpretation vertex of the Assessment Triangle is actually two things: the elicitation of bits of evidence and the accumulation of this evidence to make an inference about students’ knowledge, skills or attitudes (KSAs). Both things must be defensible, including showing accuracy and precision of the metrics involved and ruling out alternative hypotheses, as well as verifying that the assessment is fair and equitable for sub-populations. Before reporting, both the elicitation and aggregation of evidence should be transparent, justified and warranted.

### **A PRINCIPLE-DESIGNED APPROACH TO MAKE SENSE OF COMPLEX DATA: THE EVIDENCE RULES AND STATISTICAL MODELS IN ASSESSMENT**

As summarised previously in Figure 3, two components are necessary in the process of building warrants or defensible interpretations in large-scale assessments: evidence rules and the statistical model. These specify how to assign values to observable variables and how to summarise the data into indicators or scales.

#### **Building evidence rules**

Evidence rules associate a score to student actions and behaviours. Formulating such rules is rather straightforward in traditional and non-interactive assessments, particularly when multiple-choice items are used: if a student selects a correct answer, then they receive credit. More complex performance tasks require assessment designers to describe the characteristics of work products or other tangible

evidence that domain experts would associate with the KSAs in the domain of interest. In simulation- or game-based assessments, evidence rules often rely on interpreting actions and behaviours that are recorded as process data (see Box 9 for an example).


However, the interpretation of process data is susceptible to error as actions in open and interactive digital environments can often be interpreted in different ways. For example, observing that a test taker interacts with all the affordances of a simulation environment could be interpreted as demonstrating high engagement (i.e. the student confidently explores possibilities) or, conversely, high disengagement (i.e. the student does not engage meaningfully with the task). Defining evidence rules in these environments therefore requires: (1) reconstructing the universe of possible actions that the test taker can take and classifying them into meaningful groups; (2) defining the extent to which actions depend on the state of the environment (and thus on previous actions); and (3) using this information to identify sequences of contextualised actions that demonstrate mastery of the target KSAs and that can be transformed into descriptive indicators or scores.

**BOX 9.**  
**USING PROCESS DATA AS SOURCES OF EVIDENCE**

**The case of the “I like that” unit in the PISA 2025 Learning in the Digital World (LDW) assessment**

In a prototype task for the PISA LDW assessment designed to elicit evidence on students’ ability to ‘conduct experiments and analyse data’ (image below), students must use an experimentation tool to conduct experiments in which they use the control of variables strategy (CVS, i.e. varying the values of the independent variable while keeping all other variables constant).

I like that! Example →

 Complete the model.

- Conduct **experiments** to find out how ticket price impacts movie rating
- Select the **graph** that matches your results
- Select which **experiments** support your selection

**Experiments**

Experiment n.	Distance of Cinema	Ticket Price	Movie Rating
1			9
2			8
3			7
4			

**Add Experiment**



**Model** **Check work**

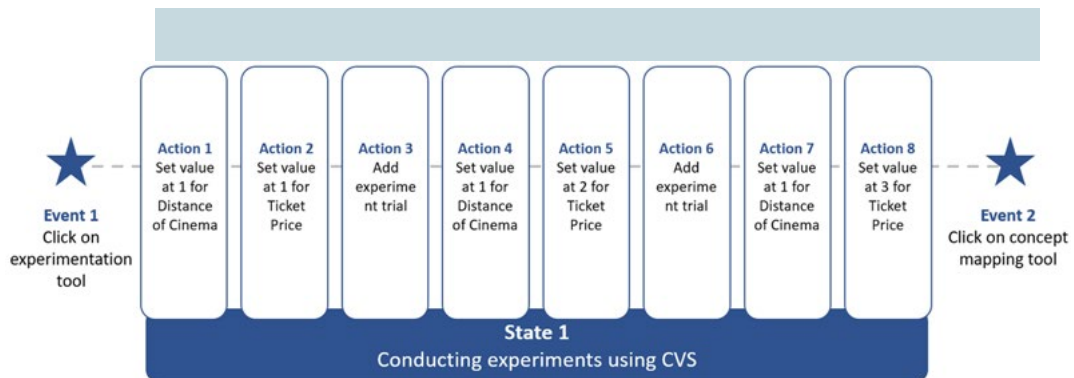
Characteristics:

Release Date

Cinema Distance

Friends' Reviews


+
→




Sequence of actions for implementing control of variable strategy

To assign a score to a student's work, sequences of actions captured in the log data are compared to an expert solution (image above). Partial credit rules can be developed to recognise students whose process data reveal they have understood the logic of controlled experiments but who made some procedural mistake in executing the strategy (e.g. testing only few values of the independent variable). Like other similar technology-enhance tasks, it is important to consider the threat of construct-irrelevant variance when defining evidence rules. One example of construct-irrelevant variance in this prototype task could be the inability of the student to conduct CVS (or any experiment at all) because they are unable to use the drop-down menus of the experimentation tool.

**Source:** OECD (forthcoming), PISA 2025 Learning in the Digital World Assessment Framework (first draft).

In the process of defining evidence rules for complex assessments, designers frequently have to revise their task designs either to add affordances to capture targeted actions or to make the environment more constrained to reduce the range of possible actions and interpretations. An iterative cycle of empirical analyses and discussions with subject-matter experts is therefore essential for evidence identification in interactive environments. This process often combines *a priori* hypotheses about the relationships between observables and KSAs with exploratory data analysis and data.

Mislevy et al. (2012) describe this interplay between theory and discovery for an assessment activity involving the configuration of a computer network. The researchers ran confirmatory analysis on a set of scoring rules defined by experts, which considered characteristics of test takers' submitted work products (for example, a given section of the network is considered 'correct' if data transfer from one computer to another). They complemented this evidence from work products by applying data mining methods to time-stamped log-file

entries. This analysis identified certain features, including the number of commands used to configure the network, the total time taken, and the number of times that students switched between networking devices, as additional potential evidence that could be combined into a measure of efficiency.

### **Selecting an appropriate statistical model**

The second component of the interpretation vertex is the statistical model that summarises data across tasks or assessment situations, in terms of updated beliefs about student-model variables. The objective in the statistical model is to express, in probabilistic terms, the relationship between observed variables (responses, final work products, sequences of actions) and the students' KSAs. Modelling specifications described in the assessment framework provide a basis for operational decisions during test construction such as deciding how many tasks are needed to make defensible conclusions based on test scores.

The simplest measurement models sum correct responses to make conclusions on proficiency, whereas more complex measurement models use latent variable frameworks such as item response (e.g., de Ayala, 2009; Reckase, 2009), diagnostic classification models (e.g. Rupp, Templin and Henson, 2010), and Bayesian networks (e.g. Levy and Mislevy, 2004; Conati, 2002).

Innovative assessments that simulate open learning and problem solving can generate evidence on students' capabilities that has high value, but that is more challenging to accumulate in existing measurement models. Because the structure and nature of data collected in technology-rich tasks can vary widely across examinees, and because test items might effectively become interdependent in open and extended tasks, it makes it difficult or inappropriate to apply the same psychometric methods used for more traditional assessments (Quellmalz et al., 2012). This evidence can only be fully exploited using new computational psychometric techniques. An important challenge ahead for innovating assessments is to refine and harness the potential of computational methods for dealing with the richer data from open and interactive environments, while preserving the inferential strengths of established psychometric methods.

### **A TALE OF TWO WORLDS: MACHINE-LEARNING APPROACHES AND EVIDENCE-CENTERED DESIGN**

Scholars in learning analytics (LA) and educational data-mining (EDM) have made tremendous progress in applying machine-learning (ML) techniques to glean useful insights from the streams of data generated in open, digital learning environments. The goal of this research is often to describe how learners learn or to find ways to adapt and personalise content to individual learners. These new methods and the rapid advances in computing technology that support them have given us the tools to identify patterns in students' thinking, even at a large scale. Assessment designers now have to take advantage of these new data-driven computational algorithms to establish new analytical models for making measurement claims, while preserving a good alignment to fundamental concepts of psychometrics.

This is not as easy as it might seem because the two fields of psychometrics and learning analytics have followed quite distinct research trajectories. Over six decades of research in psychometrics and measurement technology has established well-accepted procedures for important issues for summative assessment, which include calibration and estimation of overall score(s), reliability and precision information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data-model fit, differential functioning, and invariance. Black-box machine learning models, such as deep learning neural networks, cannot rely on such procedures, and so they are more difficult to trust when it comes to making claims about students' skills – particularly so when these claims have high stakes. Without the ability to calibrate and estimate overall scores, generate reliability and precision information, conduct sub-group analyses, and engage in linking and equating, are robust inferences from these ML models possible?

At a first glance, evidence-centered assessment design and educational data mining seem to be in conflict: the first refers to a principled approach for designing task situations that evoke particular kinds of evidence to be scored and accumulated, while the second method focuses on discovering meaningful patterns in available data. However, it is possible to use ML methods within a principled assessment design process, whereby ML models generate additional information on test takers that can be linked to evidence rules and 'aggregated' to other evidence (e.g. responses to multiple-choice items) in order to make more fine-grained and robust claims.

The simple but powerful idea behind this approach is that statistical methods that have well-established measurement properties, such as Item Response Theory (IRT), can be extended with techniques from learning analytics to fully exploit the richness of the data available in technology-rich tasks. The resulting, aggregated evidence can be evaluated using standard diagnostic procedures and can thus be more easily 'trusted' by the users of the assessment. An example of this method using a mIRT-Bayes model (Scalise, 2017) is presented in Box 10. mIRT-Bayes employs small Bayesian networks to help generate scores from patterns of actions, then uses a multidimensional IRT model to accumulate scores and yield inferences.

These opportunities for building strength across different disciplines are evident in the context of authentic technology tasks, such as simulations or serious games. Such activities incorporate many small experiences generating data patterns that are often meaningful in terms of the assessment claims. For example, an avatar controlled by the student might end up in a room with two doors; the student has then to decide which door to open and what to do in the next room. These choices can be linked to a model of students' traits and skills, and so can be used as evidence to update beliefs about students' mastery of these traits and skills. The way forward is to develop a measurement framework that encompasses perspectives from both disciplines and that supports the design and analysis of both traditional and innovative assessments (Mislevy et al., 2012).

BOX 10.

### APPLYING HYBRID MODELS TO THE TASK

#### “There’s a New Frog in Town”

The New Frog VPA is an immersive virtual environment with the look and feel of a videogame. Each participant engages as an avatar that can move around the virtual environment. The reporting goals of the assessment were multidimensional, and involved scientific exploration and inquiry (as reflected in science standards at the time).



An example screen of The New Frog VPA

In New Frog, examines were asked to explore the problem of a frog with six legs. They could choose to examine different frogs to investigate the problem, whereby the choice in itself was neither right nor wrong (so, this was not a typical ‘item’ with a pre-defined answer). However, patterns over the type and number of frogs examined (e.g. those located at different farms, along with water samples from the farms) was deemed construct salient information, and these patterns could be represented in a small but informative Bayes’ net.

The Bayes net accumulation added considerable information to the IRT model, showed acceptable fit to the patterns of the naturalistic task, and resulted in a reduction of the standard error of measurement (Scalise and Clarke-Midura, 2018). In fact, the scores generated by the two Bayes subnets proved to be among the three most informative ‘items’ in the task, in terms of the model’s fit in the study, despite being designed from data that was originally discarded. This is not terribly surprising given that the score was a pattern over salient observations, but the other most informative item was a significantly more expensive human-rated constructed-response item. Overall, a finer grain-size of inference was made possible on the task without additional testing time or scoring resources, and the strengths of low performing students in conducting inquiry were more evident.

**Source:** Scalise, Malcom and Kaylor (2023), Chapter 8 in *Innovating Assessments*.

## THE RATIONALE FOR MORE COMPLEX TASKS AND PRACTICAL WAYS TO USE THEM IN REPORTING

Designing the types of authentic tasks modelled from real learning and problem solving environments described throughout the *Innovating Assessments* publication is a central part of establishing the validity argument for next-generation assessments that target 21st century competencies, such as collaborative problem solving, that are essentially defined by processes.

Including more complex and authentic tasks in assessment also plays an important “signifying” role. 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 local, national, and international assessments. What is assessed will often end up being the focus of instruction; it is therefore critical that assessments represent the forms of knowledge and competency and the kinds of learning experiences we want to give more space to in classrooms. If students are expected to achieve the complex, multidimensional proficiencies needed for the worlds of today and tomorrow, then they should be able to demonstrate their proficiency. Embedding agency and relevancy in assessments is also likely to increase students’ engagement, and thus the likelihood of observing what students can do at the best of their capacity.

Many actors in assessment still feel poised on the precipice of what designing and including more authentic tasks imply for their work practices. Costs, versioning, compatibility with assessment delivery platforms and other practical considerations exist, especially in the context of large-scale assessments that are intended to be replicable and comparable. These constraints often discourage the creation of complex tasks to a few prototypes. Even when investments in designing complex tasks are made, the difficulty in applying standard measurement approaches with more complex data, as described above, often results in shortcuts that greatly reduce the value of integrating authentic and open tasks in the first place. For example, process data might be collected by the delivery platform but then not used in the evidence model, with only the final response getting coded as correct or incorrect and providing information about student proficiency.

In order for the measurement community to find practical entry points for including this more complex type of evidence, educational assessments (at least for now) may need to include a mix of newer and older item and task types and investigate how the evidence produced by different types of task formats and experiences triangulate. Such triangulation might help develop a shared understanding of the value of innovative tasks for inferences, and at the same time, make these inferences more defensible and ‘trusted’ by various stakeholders.

Another promising way forward consists of using different methods for different types of claims. For example, established measurement models might be used to build a scale that describes, in a reliable and comparable way, what problems students are able to solve. Learning analytics methods might then be used to provide more descriptive diagnostics of strategies and processes that students fol-

low on the tasks to achieve an output. This might be done through a cluster analysis that describes different ‘types’ of problem solvers, for instance. Descriptions of students’ work in each different cluster can be potentially very useful for teachers and students and provide tangible illustrations of how 21st century competencies are used across instructionally relevant contexts.



# INNOVATING ASSESSMENTS: THE ROAD AHEAD

*Innovating Assessments* reveals progress that has been made in conceptualising and operationalising critical aspects of 'next-generation assessments'. It provides a vision of what these should focus on, what they might look like, and how they should function. As such we have the beginnings of a map of the terrain we need to move through to get there and some destinations along the way. The map includes the constructs of interest, the innovations and practices needed to make progress, as well as many of the conceptual and technical obstacles to overcome to bring the vision of next-generation innovative assessment into being.

## INVESTING IN NEXT-GENERATION ASSESSMENTS

A journey of the type envisioned by *Innovating Assessments* cannot be undertaken nor will it succeed without an investment of multiple forms of capital. In the discussion that follows, three particular forms of capital are considered together with an explanation of their relevance. They include intellectual capital, fiscal capital, and political capital. Each is necessary but insufficient on its own. Collectively, they provide the capital needed to advance the theory and practice of educational assessment and maximise its societal benefit in the 21st century.

## INTELLECTUAL CAPITAL

When considering assessment innovation, no single discipline or area of expertise will be sufficient to accomplish what needs to be done. Advances to date reveal that next-generation assessment development is inherently a multidisciplinary enterprise. Different communities of experts need to work together collaboratively to help find solutions to the many conceptual and technical challenges already noted and those yet to be uncovered. Enlisting creative people from multiple backgrounds and perspectives to the enterprise of assessment design and use, and facilitating collaboration among them, is critical. Synergies need to be fostered between assessment designers, technology developers, learning scientists, domain experts, measurement experts, data scientists, educational practitioners, and policy makers.

Given that learning is embedded within social contexts and is characteristically shaped by cultural norms and expectations, we can expect performance to vary across cultures. Designing valid assessments, particularly those for complex skills for which established learning progressions are not available, requires multidisciplinary teams and expertise. Therefore, it is necessary to consider the complex sociocultural context in deciding what to assess, how to assess it, and how assessment results will be interpreted and used. The PISA 2022 Creative Thinking Assessment (OECD, 2020) exemplifies the importance of considering threats to comparability across languages and cultural groups when designing an assessment of a complex construct.

In addition to design and validation concerns arising from context and culture, the assessment development community writ large will need to grapple with complex issues including designing tasks that can simulate authentic contexts and elicit relevant behaviours and evidence, how to interpret and accumulate the numerous sources of data that technology enhanced assessments can generate, and how to compare students meaningfully in increasingly dynamic and open test environments. To address these and related issues considerable research will need to focus on modelling and validating complex technology enabled performances that yield multifaceted data sets. This includes modelling dependencies and non-random missing data in open and extended assessment tasks.

Emerging studies have shown that machine learning and AI techniques can help researchers better understand and model learning processes (Kleinman et al., 2022) and can assist content experts in efficiently and effectively annotating students' entire problem-solving processes at scale (Guo et al., 2022). Work of this type is needed to supplement evidence derived from small-scale cognitive lab studies and advance learning science.

At a pragmatic level, Schwartz and Arena (2013) argue that we need to 'democratise' assessment design, in the same way the design of videogames has become more accessible with the proliferation of online communities. Crowdsourcing platforms, such as the [PILA](#) system at the OECD (OECD, 2023), provide developers with model tasks they can iterate, and embed data collection instruments that simplify researchers' work on validation and measurement. Such environments and testbeds could make it far easier to engage in some of the multidisciplinary intellectual work noted above.

In summary, there are multiple intellectual and pragmatic challenges in merging learning science, data science and measurement science to understand how the sources of evidence we can obtain from complex tasks can best be analysed and interpreted using models and methods from artificial intelligence, machine learning, statistics, and psychometrics. Collaborative engagement with these concerns by learning scientists, data scientists, measurement experts, assessment designers, technology experts, and educational practitioners could yield a new discipline of Learning Assessment Engineering.

## FISCAL CAPITAL

Development of assessments for application and use at any reasonable level of scale is a time consuming and costly enterprise. The bulk of the substantial funds currently expended at national and international levels on assessment programmes is for the design and execution of large-scale assessments focused on traditional disciplinary domains like mathematics, literacy, and science (e.g. the NAEP programme in the United States and the OECD's PISA programme). Most such assessments fall within conventional parameters for task development, delivery, data capture, scoring and reporting. This has been true for quite some time despite the fact that most large-scale assessment programmes have moved to technology-based task presentation, data capture and reporting. Capitalising on many of the affordances of technology as described earlier has not been a distinct feature of those assessment programmes.

Developing and validating technology-rich tasks and environments is a much more costly activity than updating current assessments by generating traditional items using standard task designs and specifications and presenting them via technology rather than paper and pencil. Such new instruments require considerable research and development regarding task design, implementation, data analysis, scoring, reporting, and validation. As noted above, that scope of work needs to be executed by interdisciplinary groups representing domain experts, problem developers, psychometricians, UI designers and programmers. Sustained funding for the type of research and development needed is a key element in advancing next generation assessment.

A significant roadblock to achieving assessment of 21st century knowledge and skills is the paucity of examples of assessment instruments of complex cognitive construct, especially examples that have been built following systematic design principles such as Evidence-Centered Design and then validated in the field. Those cases where the work has advanced to the point where validity arguments can be offered, including evidence of feasibility for implementation at scale, have seldom moved beyond the research and development labs where they were prototyped. This is true even for cases that have achieved a high level of visibility within the assessment research and development technical community. Regrettably, this body of work has not managed to change the way assessment is conceptualised and executed at scale.

Of equal need is investment in bringing existing innovative assessments efforts to full maturity by scaling up their implementation when evidence exists that they can effectively address the challenge of measuring the constructs that matter. Current and future innovative assessment solutions are likely to languish within the R&D laboratory unless funding can be provided to move them out of the laboratory and into the space of large-scale implementation where their efficacy and utility can be properly evaluated. Only then will the possibility exist of using them to replace current ways of doing business.

## POLITICAL CAPITAL

As currently practiced educational assessment is a highly entrenched enterprise, particularly the use of large-scale standardised assessments for educational monitoring and policy decisions. Standardisation includes what is assessed, how it is assessed, how the data are collected and then analysed, and how the results are interpreted and then reported. This is not an accident but the product of many years of operating within a particular perspective on what we want and need to know about the knowledge, skills and abilities of individuals coupled with a highly refined technology of test development and administration that is further coupled with an epistemology of interpretation about the mental world rooted in a measurement metaphor derived from the physical world.

It is hard to make major changes within existing systems when there are well established operational programmes 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. 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 achieve meaningful change in educational assessment.

The political capital needed is not limited to state and federal policymakers. It encompasses multiple segments of the educational assessment development community, the measurement and psychometric community, and the educational practice community. Each of these communities has entrenched assumptions and practices when it comes to assessment. Thus, each community needs to buy into a vision of transformation that may well yield outcomes at variance with aspects of current standard operating procedure. For example, if a student's knowledge and skills are seen as no longer discrete and independent then assessing them may require examining the entire interactive behaviour/process in adaptive learning environments that mimic real-world scenarios. Regardless of where the process may lead, these communities must work together to generate the amount of political will and capital needed to organise, support, and sustain such process.

## **INTERNATIONAL LARGE-SCALE ASSESSMENTS: POSSIBILITIES FOR INNOVATION AT SCALE**

It should be obvious that much is needed to advance the agenda for innovation in assessment along the lines previously outlined. One of the biggest challenges in making change happen is that scale is needed to show what is possible. Scaling up promising ideas is critical for testing how flexible or brittle those ideas and assessment approaches may be, in addition to what it takes to put them into practice at scale. Fortunately, we have some examples of efforts to do so, which teach us much with respect to what is possible, as well as where challenges remain.

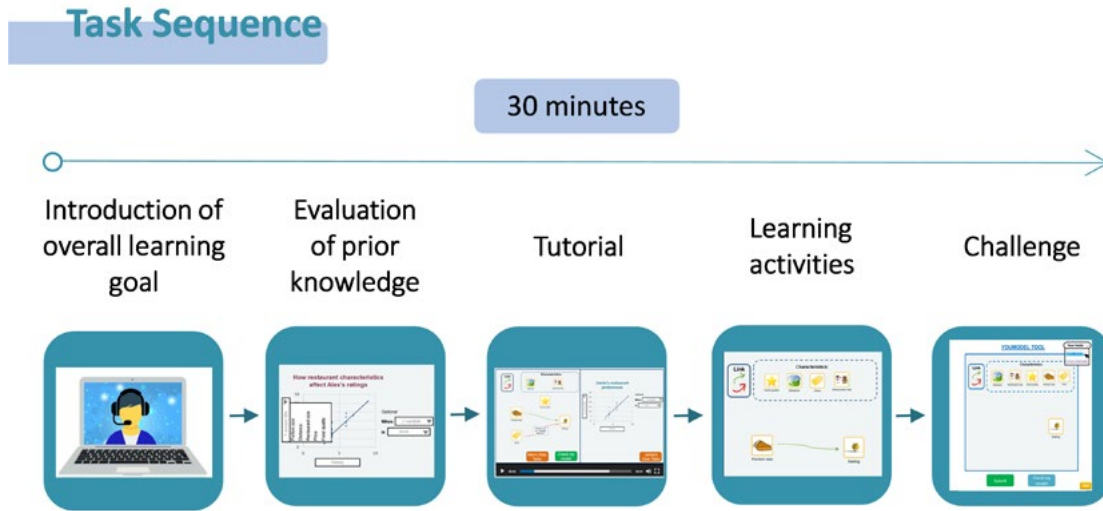
International assessments generally serve as tools for monitoring performance on contemporary disciplinary standards. As such these programmes make statements about what is valued globally and provide information about student proficiency at scale. They also illustrate an operational example of the pooling of intellectual, fiscal, and political capital required to move an innovative, large-scale assessment agenda forward. For example, in addition to its ongoing regular assessment programmes in mathematics, reading and science, OECD's PISA Programme has embarked on including one "innovative" assessment in each of its assessment cycles. Through this effort, the OECD has signalled the important forms of 21st century knowledge and skill that should be assessed as a part of monitoring broader educational goals and aims. We will briefly consider one recent example from that programme to illustrate some of what has been learned through attempts to put innovative ideas about the assessment of learning into practice.

### **PISA 2025 LEARNING IN THE DIGITAL WORLD**

In its 2025 cycle, PISA will include an assessment of Learning in the Digital World. When the PISA Governing Board embarked on this new development back in 2020, there were clear expectations about the added value it should bring: countries were interested in comparable data on students' readiness to learn and problem solve with digital tools. Even before the COVID-19 global pandemic, it was clear to stakeholders that digital technologies are significantly impacting education, yet there is not enough information on whether students have the necessary skills to learn with these new tools and on whether schools are equipped to support these new ways of learning.

This policy demand oriented several design decisions. As already discussed, an assessment of learning skills has different requirements from an assessment of knowledge. To distinguish more effective learners from less effective learners, the assessment had to provide opportunities for students to engage in some type of knowledge construction activities. In other words, the assessment designers had to structure the assessment as a learning experience, where it would be possible to evaluate how students' knowledge changed over the course of the assessment. Consequently, the structure of the assessment units has diverged from the traditional PISA format, with a series of stimuli and independent questions, to a new format that is structured as a series of connected lessons (Figure 6).

**Figure 6.** Task sequence in the PISA 2025 Learning in the Digital World assessment



**Source:** OECD (forthcoming), PISA 2025 Learning in the Digital World assessment framework (draft), OECD Publishing, Paris.

A virtual tutor guides the students through the test, explaining how they can solve relatively complex problems using digital tools that include block-based coding, simulations, data collection and modelling interfaces. An interactive tutorial with videos is embedded in each unit to help students understand how to use these tools and mitigate differences in students' familiarity with particular digital tools or learning environments. Students then solve a series of tasks that progress from easier to more difficult, introducing them to the concepts and practices they are expected to learn in the unit and that they will need to apply to the final and more complex "challenge" task.

Part of the assessment construct relates to students' capacity to engage in self-regulated learning, therefore requiring the development of measures such as monitoring and adapting to feedback, and evaluating knowledge and performance. In order to generate observables for these self-regulated learning processes, a number of affordances were embedded in the assessment environment. Over the course of the test, students can receive feedback by testing whether they achieve the expected outcomes by asking the tutor to check their work. They can choose to see the solutions to the training tasks after they submit their answers, and for each task they can access hints and worked examples to help them solve the problem. At the end of each challenge task, students are asked to evaluate their performance and report the effort they invested while working through the unit and the emotions they felt during 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 solve problems, but how they go about learning to do so.

These innovations represent responses to well-defined evidentiary needs. The assessment has been designed to provide responses to three interconnected questions: what types of problems in the domain of computational design and modelling can students solve? To what extent are they able to learn new concepts in this domain by solving sequences of connected, scaffolded tasks? And to what extent is this learning supported by productive behaviours, such as decisions to use learning affordances when needed or monitor progress towards their learning goals? These questions have defined the cognition model of the assessment, have oriented the design of tasks needed to elicit the necessary observations, and are guiding analysis plans to interpret the data in a way that is consistent with the reporting purposes of the assessment and that accounts for the complex nature of the data.

The expectation is to produce multi-dimensional reports of student performance on this test, including measures of students' (1) overall performance on the tasks (represented in a scale, as in other PISA assessments); (2) learning gains, i.e. how much students' knowledge of given concepts and their capacity to complete specific operations increases following the training; and (3) capacity to self-regulate their learning and manage their affective states. These different measures will be triangulated in the analysis, for example, explaining part of the variation in learning gains with the indicators of self-regulated learning behaviours. The goal is to provide policy-makers with actionable information that is not limited to one score and a position in an international ranking, but that includes more nuanced descriptions of what students can do and indicates what aspects of their performance deserve more attention.

## CODA: RETURNING TO THE THREE TYPES OF CAPITAL

The development of the PISA 2025 Learning in the Digital World assessment was only possible because of the convergence of the different types of capital described above. The political backing of a research and development agenda by PISA participating countries has been strong. 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 then be transferred to the trend domains of reading, mathematics and science or that can provide inspiration for the development of national assessments once their value is proven.

Acknowledging the need for multiple iterations in the design of tasks and for extensive validation processes for design and analytical choices through cognitive laboratories and pilot studies, the PISA Governing Board provided the financial and political support needed to start the development of the test five years before the main data collection. Further resources were made available by research foundations that recognised 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 groups.

This new PISA test represents only an initial foray into the enterprise of innovating assessments. As argued in *Innovating Assessments*, we need many new disciplinary and cross-disciplinary assessments to provide an exhaustive description of the quality of educational experiences across countries. Several challenges also remain, particularly in the interpretation vertex of the Assessment Triangle. International fora, like PISA or the IEA, have a role to play in coordinating policy demands and facilitating a consensus on what pieces of the puzzle we need to work on and what the priorities should be for the near term and beyond. There is more than ample evidence that innovative assessment of educationally and socially significant competencies is both desirable and possible. The evidence also suggests that cooperation and collaboration on a global scale may well be the best and only way to achieve such advances.



# REFERENCES

- Aleven, V. et al. (2016), "Help Helps, But Only So Much: Research on Help Seeking with Intelligent Tutoring Systems", *International Journal of Artificial Intelligence in Education*, Vol. 26/1, pp. 205-223, <https://doi.org/10.1007/s40593-015-0089-1>.
- Baines, E., P. Blatchford and A. Chowne (2007), "Improving the effectiveness of collaborative group work in primary schools: Effects on science attainment", *British Educational Research Journal*, Vol. 33/5, pp. 663-680, <https://doi.org/10.1080/01411920701582231>.
- Basol, M. et al. (2021), "Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation", *Big Data & Society*, Vol. 8/1, p. 205395172110138, <https://doi.org/10.1177/20539517211013868>.
- Bellanca, J. (2014), *Deeper learning: Beyond 21st century skills*, Solution Tree Press, Bloomington.
- Bilal, D. (2000), "Children's use of the Yahoo!igans! web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks", *Journal of the American Society for Information Science*, Vol. 51/7, pp. 646-665, [https://doi.org/10.1002/\(SICI\)1097-4571\(2000\)51:73.0.CO;2-A](https://doi.org/10.1002/(SICI)1097-4571(2000)51:73.0.CO;2-A).
- Binkley, M. et al. (2011), "Defining Twenty-First Century Skills", in *Assessment and Teaching of 21st Century Skills*, Springer Netherlands, Dordrecht, [https://doi.org/10.1007/978-94-007-2324-5\\_2](https://doi.org/10.1007/978-94-007-2324-5_2).
- Biswas, G., J. Segedy and K. Bunchongchit (2015), "From design to implementation to practice a learning by teaching system: Betty's Brain", *International Journal of Artificial Intelligence in Education*, Vol. 26/1, pp. 350-364, <https://doi.org/10.1007/s40593-015-0057-9>.
- Brand-Gruwel, S., I. Wopereis and Y. Vermetten (2005), "Information problem solving by experts and novices: Analysis of a complex cognitive skill", *Computers in Human Behavior*, Vol. 21/3, pp. 487-508, <https://doi.org/10.1016/j.chb.2004.10.005>.
- Bransford, J. and B. Stein (1984), *The Ideal Problem Solver: A Guide for Improving Thinking, Learning, and Creativity*, Freeman, New York.
- Clark, R. et al. (2008), "Cognitive task analysis", in Spector J. et al. (eds.), *Handbook of Research on Educational Communications and Technology*, Macmillan/Gale, New York, pp. 541-551.
- Coiro, J. et al. (2019), "Students engaging in multiple-source inquiry tasks: Capturing dimensions of collaborative online inquiry and social deliberation", *Literacy Research: Theory, Method, and Practice*, Vol. 68/1, pp. 271-292, <https://doi.org/10.1177/2381336919870285>.
- Conati, C. (2002), "Probabilistic assessment of user's emotions in educational games", *Applied Artificial Intelligence*, Vol. 16/7-8, pp. 555-575, <https://doi.org/10.1080/08839510290030390>.

- de Ayala, R. (2009), *The Theory and Practice of Item Response Theory*, Guilford Press.
- Jong, T. et al. (2018), "Simulations, Games, and Modeling Tools for Learning", in *International Handbook of the Learning Sciences*, Routledge, New York, NY : Routledge, 2018., pp. 256-266, <https://doi.org/10.4324/9781315617572-25>.
- Deeva, G. et al. (2021), "A review of automated feedback systems for learners: Classification framework, challenges and opportunities", *Computers & Education*, Vol. 162, p. 104094, <https://doi.org/10.1016/j.compedu.2020.104094>.
- Ercikan, K. and M. Oliveri (2016), "In search of validity evidence in support of the interpretation and use of assessments of complex constructs: Discussion of research on assessing 21st century skills", *Applied Measurement in Education*, Vol. 29/4, pp. 310-318, <https://doi.org/10.1080/08957347.2016.1209210>.
- Ercikan, K. and J. Pellegrino (2017), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, <https://doi.org/10.4324/9781315708591>.
- Foster, N. and M. Piacentini (eds.) (2023), *Innovating Assessments to Measure and Support Complex Skills*, OECD Publishing, Paris, <https://doi.org/10.1787/e5f3e341-en>.
- Ganaiem, E. and I. Roll (2022), "The effect of different sequences of examples and problems on learning experimental design", *Proceedings of the International Conference of the Learning Sciences*, Hiroshima, pp. 727-732.
- Gillies, R. (2016), "Cooperative learning: Review of research and practice", *Australian Journal of Teacher Education*, Vol. 41/3, pp. 39-54, <https://doi.org/10.14221/ajte.2016v41n3.3>.
- Gillies, R. and M. Boyle (2010), "Teachers' reflections on cooperative learning: Issues of implementation", *Teaching and Teacher Education*, Vol. 26/4, pp. 933-940, <https://doi.org/10.1016/j.tate.2009.10.034>.
- Glogger-Frey, I. et al. (2015), "Inventing a solution and studying a worked solution prepare differently for learning from direct instruction", *Learning and Instruction*, Vol. 39, pp. 72-87, <https://doi.org/10.1016/j.learninstruc.2015.05.001>.
- Guo, H., Johnson, M., Ercikan, K., Saldivia, L. & Worthington, M. (2022, July). Understanding students' test performance and engagement (Invited session organized/chaired by K. Ercikan). International Meeting of Psychometric Society, Bologna, Italy.
- Guzdial, M., J. Rick and C. Kehoe (2001), "Beyond adoption to invention: Teacher-created collaborative activities in higher education", *Journal of the Learning Sciences*, Vol. 10/3, pp. 265-279, [https://doi.org/10.1207/s15327809jls1003\\_2](https://doi.org/10.1207/s15327809jls1003_2).
- Hubley, A. and B. Zumbo (2017), "Response Processes in the Context of Validity: Setting the Stage", in *Understanding and Investigating Response Processes in Validation Research, Social Indicators Research Series*, Springer International Publishing, Cham, pp. 1-12, [https://doi.org/10.1007/978-3-319-56129-5\\_1](https://doi.org/10.1007/978-3-319-56129-5_1).
- Irava, V. et al. (2019), "Game-based socio-emotional skills assessment: A comparison across three cultures", *Journal of Educational Technology Systems*, Vol. 48/1, pp. 51-71, <https://doi.org/10.1177/0047239519854042>.
- Jonassen, D. (1992), "What are Cognitive Tools?", in *Cognitive Tools for Learning*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1-6, [https://doi.org/10.1007/978-3-642-77222-1\\_1](https://doi.org/10.1007/978-3-642-77222-1_1).

- Kinnebrew, J., J. Segedy and G. Biswas (2017), "Integrating Model-Driven and Data-Driven Techniques for Analyzing Learning Behaviors in Open-Ended Learning Environments", *IEEE Transactions on Learning Technologies*, Vol. 10/2, pp. 140-153, <https://doi.org/10.1109/tlt.2015.2513387>.
- Kleinman, E. et al. (2022), "Analyzing Students' Problem-Solving Sequences", *Journal of Learning Analytics*, pp. 1-23, <https://doi.org/10.18608/jla.2022.7465>.
- Large, A. and J. Beheshti (2000), "The web as a classroom resource: Reactions from the users", *Journal of the American Society for Information Science*, Vol. 51/12, pp. 1069-1080, [https://doi.org/10.1002/1097-4571\(2000\)9999:9999::AID-ASI1017>3.0.CO;2-W](https://doi.org/10.1002/1097-4571(2000)9999:9999::AID-ASI1017>3.0.CO;2-W).
- Levy, R. and R. Mislevy (2004), "Specifying and refining a measurement model for a computer-based interactive assessment", *International Journal of Testing*, Vol. 4/4, pp. 333-369, [https://doi.org/10.1207/s15327574ijt0404\\_3](https://doi.org/10.1207/s15327574ijt0404_3).
- Lubart, T. (1990), "Creativity and Cross-Cultural Variation", *International Journal of Psychology*, Vol. 25/1, pp. 39-59, <https://doi.org/10.1080/00207599008246813>.
- Messick, S. (1994), "The Interplay of Evidence and Consequences in the Validation of Performance Assessments", *Educational Researcher*, Vol. 23/2, p. 13, <https://doi.org/10.2307/1176219>.
- Mislevy, R. et al. (2012), "Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining", *Journal of Educational Data Mining*, Vol. 4/1, pp. 11-48, <https://doi.org/10.5281/zenodo.3554641>.
- Mislevy, R. and G. Haertel (2007), "Implications of Evidence-Centered Design for Educational Testing", *Educational Measurement: Issues and Practice*, Vol. 25/4, pp. 6-20, <https://doi.org/10.1111/j.1745-3992.2006.00075.x>.
- Mislevy, R. and M. Riconscente (2006), "Evidence-centered assessment design: Layers, concepts, and terminology", in Downing, S. and T. Haladyna (eds.), *Handbook of test development*, Erlbaum, Mahwah, NJ, pp. 61-90.
- Nathan, M. (1998), "Knowledge and Situational Feedback in a Learning Environment for Algebra Story Problem Solving", *Interactive Learning Environments*, Vol. 5/1, pp. 135-159, <https://doi.org/10.1080/1049482980050110>.
- Niu, W. and R. Sternberg (2001), "Cultural influences on artistic creativity and its evaluation", *International Journal of Psychology*, Vol. 36/4, pp. 225-241, <https://doi.org/10.1080/00207590143000036>.
- OECD (forthcoming), *PISA 2025 Learning in the Digital World assessment framework (draft)*, OECD Publishing, Paris
- OECD (2022), *Thinking Outside the Box: The PISA 2022 Creative Thinking Assessment*, <https://issuu.com/oecd.publishing/docs/thinking-outside-the-box> (accessed on 4 March 2023).
- OECD (2013), *Synergies for Better Learning: An International Perspective on Evaluation and Assessment*, OECD Publishing, Paris, <http://www.oecd-ilibrary.org/docserver/download/9113021e.pdf?expires=1511446761&id=id&accname=guest&checksum=18A9C-C493392BE9A918508D9929D29A3>.

Pellas, N. et al. (2018), "Augmenting the learning experience in primary and secondary school education: a systematic review of recent trends in augmented reality game-based learning", *Virtual Reality*, Vol. 23/4, pp. 329-346, <https://doi.org/10.1007/s10055-018-0347-2>.

Pellegrino, J., N. Chudowsky and R. Glaser (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academy Press.

Pellegrino, J., L. DiBello and S. Goldman (2016), "A Framework for Conceptualizing and Evaluating the Validity of Instructionally Relevant Assessments", *Educational Psychologist*, Vol. 51/1, pp. 59-81, <https://doi.org/10.1080/00461520.2016.1145550>.

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

Quellmalz, E. et al. (2012), "21st century dynamic assessment", in Mayrath, M. et al. (eds.), *Technology-Based Assessments for 21st Century Skills*, Information Age Publishing,, [http://www.simsScientists.org/downloads/Chapter\\_2012\\_Quellmalz.pdf](http://www.simsScientists.org/downloads/Chapter_2012_Quellmalz.pdf).

Raphael, C. et al. (2009), "Games for civic learning: A conceptual framework and agenda for research and design", *Games and Culture*, Vol. 5/2, pp. 199-235, <https://doi.org/10.1177/1555412009354728>.

Reckase, M. (2009), *Multidimensional Item Response Theory*, Springer, New York, <https://doi.org/10.1007/978-0-387-89976-3>.

Roll, I. et al. (2018), "Understanding the impact of guiding inquiry: the relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning", *Instructional Science*, Vol. 46/1, pp. 77-104, <https://doi.org/10.1007/s11251-017-9437-x>.

Roll, I. et al. (2014), "Tutoring Self- and Co-Regulation with Intelligent Tutoring Systems to Help Students Acquire Better Learning Skills", in Sottolare, R. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems: Volume 2 Instructional Management*, US Army Research Laboratory, Orlando, pp. 169-182.

Roozenbeek, J. and S. van der Linden (2018), "The fake news game: Actively inoculating against the risk of misinformation", *Journal of Risk Research*, Vol. 22/5, pp. 570-580, <https://doi.org/10.1080/13669877.2018.1443491>.

Rupp, A., J. Templin and R. Henson (2010), *Diagnostic Measurement: Theory, Methods, and Applications*, Guilford Press, New York.

Scalise, K. (2017), "Hybrid Measurement Models for Technology-Enhanced Assessments Through mIRT-bayes", *International Journal of Statistics and Probability*, Vol. 6/3, p. 168, <https://doi.org/10.5539/ijsp.v6n3p168>.

Scalise, K. and J. Clarke-Midura (2018), "The many faces of scientific inquiry: Effectively measuring what students do and not only what they say", *Journal of Research in Science Teaching*, Vol. 55/10, pp. 1469-1496, <https://doi.org/10.1002/tea.21464>.

Schwartz, D. and D. Arena (2013), *Measuring what matters most: Choice-based assessments for the digital age*, The MIT Press, Cambridge, Massachusetts.

Seo, K. et al. (2021), "Active learning with online video: The impact of learning context on engagement", *Computers & Education*, Vol. 165, p. 104132, <https://doi.org/10.1016/j.compedu.2021.104132>.

- Sternberg, R. (2013), "Intelligence", in Freedheim, D. and I. Weiner (eds.), *Handbook of Psychology: History of Psychology*, John Wiley & Sons, Hoboken, pp. 155-176.
- Toulmin, S. (2003), *The Uses of Argument*, Cambridge University Press, Cambridge, <https://doi.org/10.1017/cbo9780511840005>.
- Urban, A., C. Hewitt and J. Moore (2018), "Fake it to make it, media literacy, and persuasive design: Using the functional triad as a tool for investigating persuasive elements in a fake news simulator", *Proceedings of the Association for Information Science and Technology*, Vol. 55/1, pp. 915-916, <https://doi.org/10.1002/pra2.2018.14505501174>.
- van der Linden, S., J. Roozenbeek and J. Compton (2020), "Inoculating against fake news about COVID-19", *Frontiers in Psychology*, Vol. 11/566790, pp. 1-7, <https://doi.org/10.3389/fpsyg.2020.566790>.
- VanLehn, K. et al. (2007), *What's in a step? Toward general, abstract representations of tutoring system log data*, Springer.
- Voogt, J. and N. Roblin (2012), "A comparative analysis of international frameworks for 21<sup>st</sup> century competences: Implications for national curriculum policies", *Journal of Curriculum Studies*, Vol. 44/3, pp. 299-321, <https://doi.org/10.1080/00220272.2012.668938>.
- Wainer, H. et al. (2000), *Computerized Adaptive Testing*, Routledge, <https://doi.org/10.4324/9781410605931>.
- Wieman, C., W. Adams and K. Perkins (2008), "PhET: Simulations That Enhance Learning", *Science*, Vol. 322/5902, pp. 682-683, <https://doi.org/10.1126/science.1161948>.
- Winstone, N. et al. (2016), "Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes", *Educational Psychologist*, Vol. 52/1, pp. 17-37, <https://doi.org/10.1080/00461520.2016.1207538>.
- Wolf, S., T. Brush and J. Saye (2003), "Using an information problem-solving model as a metacognitive scaffold for multimedia-supported information-based problems", *Journal of Research on Technology in Education*, Vol. 35/3, pp. 321-341, <https://doi.org/10.1080/15391523.2003.10782389>.
- Wood, D. (2001), "Scaffolding, Contingent Tutoring and Computer-supported Learning", *International Journal of Artificial Intelligence in Education*, Vol. 12, pp. 280-293.







Support:

