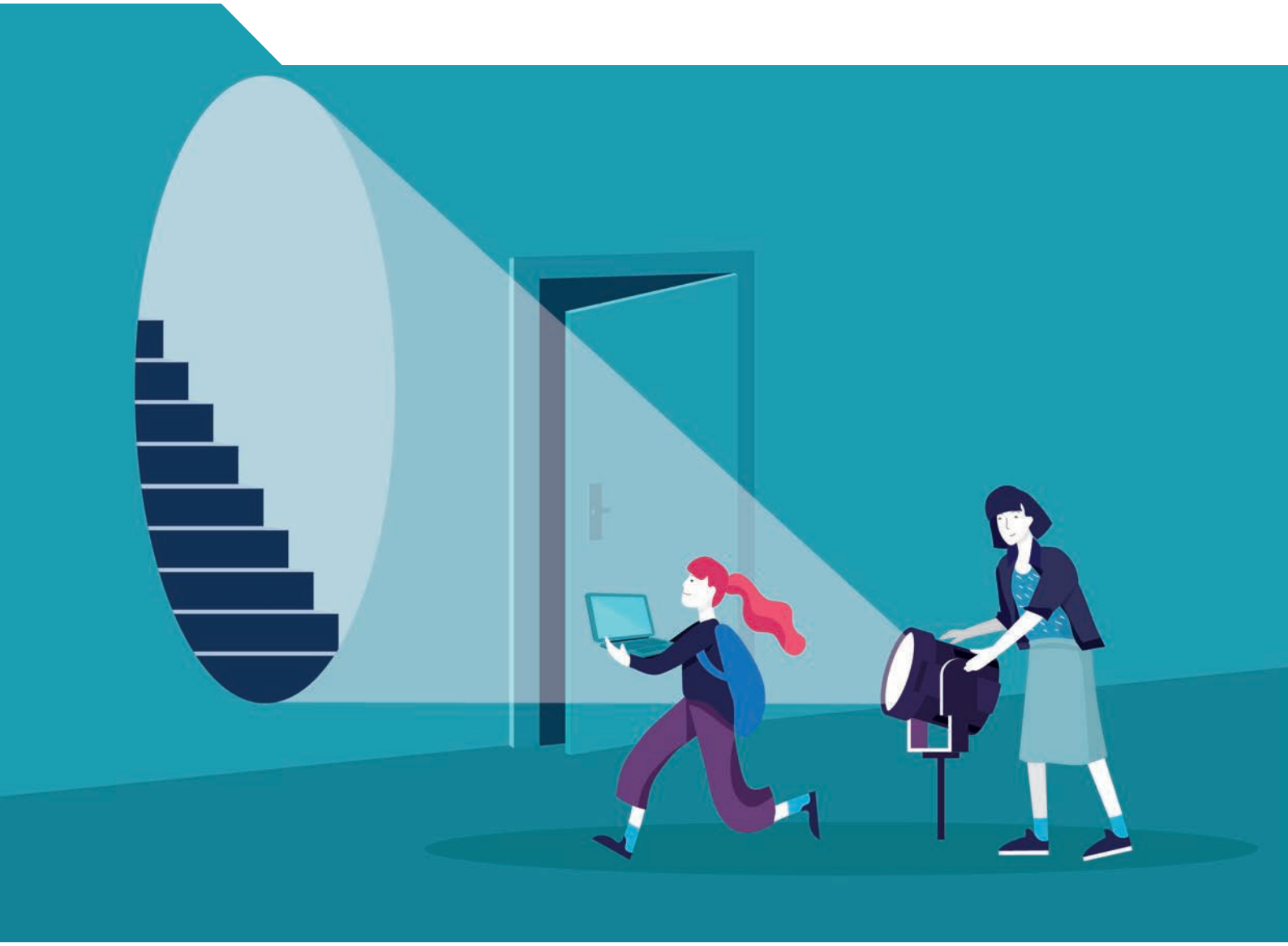
Innovating Assessments to Measure and Support Complex Skills

Edited by Natalie Foster and Mario Piacentini



OECD

Innovating Assessments to Measure and Support Complex Skills

Edited by
Natalie Foster and Mario Piacentini

OECD
BETTER POLICIES FOR BETTER LIVES

Please cite this publication as:
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.

ISBN 978-92-64-66443-2 (print)
ISBN 978-92-64-37850-6 (pdf)
ISBN 978-92-64-56338-4 (HTML)
ISBN 978-92-64-52836-9 (epub)

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

Corrigenda to OECD publications may be found on line at: www.oecd.org/about/publishing/corrigenda.htm.

Foreword

Back in 2018, while working at the OECD Secretariat, the editors of this report were developing the first global assessment of students' creative thinking. Countries around the world were also administering the latest cycle of the OECD's Programme for International Student Assessment (PISA) tests, which for the first time included an assessment of global competence. These two "innovative domains" represented new frontiers for PISA. The global competence assessment aimed to provide comparative evidence on a multidimensional construct that included attitudinal and socio-emotional elements. The creative thinking assessment was also being developed to include open-ended items across four different domain areas, with complex scoring rubrics guiding human coders. These experiences made it clear that developing new assessments of complex skills was – unsurprisingly – a highly complex task, starting from how such constructs should be defined through to how they should be measured, validated and reported.

The editors then had an idea: what if PISA could bring together an international group of senior experts in educational measurement and assessment design to discuss the next generation of educational assessments? In 2019, the PISA Research and Innovation Group was created and thus held its first meeting to discuss a strategy for the next cycles of the PISA "innovative domains", the purpose of which are to develop new measures of relevant 21st Century competencies in large-scale student assessment and to foster innovation in the way that students are assessed.

It soon became clear that the need for change in educational assessment reached beyond a single "innovative" PISA component. With that, the scope of the group's work expanded: in 2020, PISA member countries decided to establish and fund a dedicated Research, Development and Innovation programme in PISA; and the OECD began to develop the Platform for Innovative Learning Assessments (PILA), a tool for prototyping innovative assessment tasks that can be used either for formative or summative purposes.

This report is the product of a collaborative, multi-year effort between the PISA Research and Innovation Group, various experts in the field of educational measurement and assessment design, and the OECD PISA Secretariat. It summarises much of the research work carried out by members of the group since its first meeting in 2019. This report aims to provide future directions for the development of innovative PISA assessments that focus on measuring complex skills. It also aims to support countries around the world who are engaged in reforming or expanding the scope of their national assessment systems. As more countries include skills like creative thinking, self-regulation and complex problem solving in their curricula goals, it is important that assessments evolve to better measure and support these skills in ways that are valid and instructionally relevant.

This report does not aim to provide a list of new assessments that PISA or national education systems should implement. Rather, it unpacks some of the key arguments in support of innovating assessments and the challenges facing assessment designers in doing so. To this end, the report presents design principles supporting the design of assessments that are fit for their purpose, some future directions based on promising examples – and some new questions that we need to address. We hope this report inspires ambition to pursue the goal of innovating assessments, both in PISA and beyond, while highlighting areas for future research and collaboration.

Acknowledgements

The editors would first like to thank Instituto Unibanco, whose ambitious vision for innovating assessments, generous funding, and support for establishing the PISA Research and Innovation Group made the conception and development of this report possible. We also thank Bertelsmann Stiftung, Deutsche Telekom Stiftung, Robert Bosch Stiftung and Mercator Stiftung for generously supporting the development of the PISA 2025 Learning in the Digital World assessment and the PILA assessment modules that are presented in various chapters in this report.

We would also like to thank the members of the PISA Research and Innovation Group (RIG) who have guided the development of this report from the outset, drafted several chapters and provided precious peer review. RIG members include Kadriye Ercikan (Educational Testing Service), Xiangen Hu (University of Memphis), Cesar A. A. Nunes (Universidade Estadual de Campinas), James Pellegrino (University of Illinois, Chicago), Ido Roll (Technion – Israel Institute of Technology) and Kathleen Scalise (University of Oregon).

This book also benefited from the contributions of a larger number of experts and researchers. For their valuable contributions, the editors extend their gratitude to Miri Barhak-Rabinowitz (Technion – Israel Institute of Technology), Hongwen Guo (Educational Testing Service), Errol Kaylor (University of Oregon), Cassandra Malcolm (University of Oregon), Han Hui Por (Educational Testing Service), Argenta Price (Stanford University), John. P. Sabatini (University of Memphis), Keith Shubeck (University of Memphis and Carl Wieman (Stanford University). The chapter written by Carl Wieman and Argenta Price was funded by the Howard Hughes Medical Institute through a Professor grant to Carl Wieman.

Within the OECD Secretariat, this report was edited by Natalie Foster and Mario Piacentini, who both contributed several chapters to the publication. Marc Fuster Rabella also provided invaluable editorial support and comments in preparation of the final manuscript. Finally, the editors thank OECD colleagues who provided research and feedback on earlier draft chapters including Janine Buchholz, Marta Cignetti, Ava Guez and Emma Linsenmayer. We also thank Andreas Schleicher, Director for Education and Skills, and Yuri Belfali, Head of the Early Childhood and Schools Division, for their review and feedback. Sophie Limoges, Cassandra Morley, Alexandra Selee and Della Shin contributed to the final stages of preparation for publication.

Table of contents

Foreword 3

Acknowledgements 5

Editorial 11

Executive summary 13

Introduction: Arguments in support of innovating assessments 15
- Overview 16
- Assessment as a process of reasoning from evidence 16
- Argument 1: Measuring what matters 18
- Argument 2: Assessment design processes and applications of technology 19
- Argument 3: Valid interpretation and use of results 22
- Towards more coherent systems of assessment 23
- Conclusion 24
- References 25

Part I Innovating What We Assess 29

1 21st Century competencies: Challenges in education and assessment 30
- Introduction 31
- What are 21st Century competencies? 32
- A multi-faceted challenge for education systems 33
- Conclusion 40
- References 42

2 Next-generation assessments of 21st Century competencies: Insights from the learning sciences 45
- Introduction 46
- Research insights on deeper learning 46
- Design innovations for new assessments of 21st Century competencies 48
- Conclusion 56
- References 58

3 Framing the focus of new assessments of 21st Century competencies 61
- Introduction 62
- Towards a more comprehensive system of assessments: Framing initial design decisions 62

Conclusion 74
References 76

4 Assessing complex problem-solving skills through the lens of decision making 79
Introduction: Scientific problem solving as a core competency 80
A need for specificity in defining complex problem solving in S&E 80
The problem-solving process of experts in science, engineering and medicine 82
Application of the decision framework to assessment across educational levels and contexts 85
Designing tasks for decision-based assessments 88
References 93

Part II Innovating How We Assess 97

5 Exploiting technology to innovate assessment 98
Introduction 99
The unexploited potential of technology to transform assessment 99
Leveraging technology across an evidence-centred assessment design process 100
Discussion 104
Conclusion 105
References 107

6 Defining the conceptual assessment framework for complex competencies 111
Introduction 112
Complexity breeds complexity: The importance of theory to orient initial design decisions 112
Establishing solid conceptual foundations: Domain analysis and modelling 114
The three models of an ECD assessment framework 116
An illustrative example: The PISA 2025 Learning in the Digital World assessment 120
Conclusion 125
References 127

7 Designing innovative tasks and test environments 131
Introduction 132
Using technology to enhance task design 133
Conclusion 143
References 145

8 Analysing and integrating new sources of data reliably in innovative assessments 147
Introduction 148
Do we need new analytic approaches? 148
Some possible solutions 149
Understanding innovation challenges from a measurement perspective 150
Conclusion 157
References 158

9 Measuring self-regulated learning using feedback and resources 159
Introduction 160
Resources and affordances that support the assessment of SRL 160
Tracking evidence for SRL using resources: The PISA 2025 Learning in the Digital World assessment 162
Designing tasks and resources for SRL assessments 164

Key challenges 165
Conclusion 167
References 168

10 Artificial Intelligence-enabled adaptive assessments with Intelligent Tutors 173
Introduction 174
Why are ITS relevant to next-generation assessments? 174
An ITS adaptive assessment framework 175
Selected examples 179
Conclusion and final thoughts 184
References 186

Part III Innovating How We Interpret and Use Assessment Results 189

11 Cross-cultural validity and comparability in assessments of complex constructs 190
Introduction 191
Sociocultural context of assessment 191
Considerations for cross-cultural validity and comparability in assessments 193
Construct equivalence 194
Test equivalence 195
Test condition equivalence 197
Integrating a sociocultural perspective in assessment design 198
Methodologies for examining equivalence 200
Conclusion 201
References 203

12 Uses of process data in advancing the practice and science of technology-rich assessments 211
Introduction 212
Construction of response process indicators 213
Use of process data for improving assessment design and validity 215
Use of process data as evidence of the construct for augmenting score creation 217
Use of process data for group comparisons and fairness research 218
Conclusion 220
References 222

13 A tale of two worlds: Machine learning approaches at the intersection with educational measurement 229
Introduction 230
What is at stake? 235
Conclusion 236
References 237

14 Conclusions and implications 239
Introduction 240
Part I. Arguments and evidence regarding innovation in educational assessment 240
Part II. Innovative assessments: Progress made and the road ahead 245
References 251

FIGURES

Figure A. The Assessment Triangle 17
Figure 1.1. Broad categories of 21st Century competencies 32
Figure 2.1. Contrasting cases example for designing a zoo exhibit 52
Figure 2.2. A low floor, high ceiling task in the PILA Karel application 55
Figure 2.3. Assessment experience “map” (sequence of tasks) in the PILA Karel application 56
Figure 3.1. Framing the focus of next-generation assessments 63
Figure 3.2. Screenshot from the SimCityEdu interface 65
Figure 3.3. Screenshot from the Betty’s Brain interface 68
Figure 3.4. Asymmetric distribution of tools and information in the “Olive Oil” task 72
Figure 3.5. Example task interface for the PISA 2015 Collaborative Problem-Solving Assessment 73
Figure 4.1. *School* problems and *authentic* problems 81
Figure 4.2. “Running in Hot Weather” example problem from the PISA 2015 Scientific Literacy assessment 87
Figure 4.3. Decisions-based assessment template 89
Figure 4.4. Examples of simulations used for assessing problem solving in S&E 90
Figure 6.1. Phases of defining the conceptual assessment framework in an ECD process 114
Figure 6.2. Assessment design as a process of argumentation 116
Figure 6.3. An example of a student model for system thinking 117
Figure 6.4. Student model for the PISA 2025 LDW assessment 123
Figure 7.1. Technology-enhanced assessment (TEA) design framework 133
Figure 7.2. Task format continuum in technology-enhanced assessment design 134
Figure 7.3. Test features in technology-enhanced assessment design 137
Figure 8.1. Screenshot from original version of PILA Karel task 151
Figure 8.2. Conceptual map of original version of PILA Karel task 152
Figure 8.3. An example screen of The New Frog VPA 153
Figure 8.4. One example theoretical Bayes subnet for accumulating information from original PILA Karel task example 156
Figure 9.1. Resources in the “I Like It!” task of the PISA 2025 LDW assessment 163
Figure 10.1. Elements of an ITS adaptive assessment framework 175
Figure 10.2. The interaction between learners and avatars (tutors) in ITS 178
Figure 10.3. ElectronixTutor interface 180
Figure 10.4. The conversation flow of an Expectation-Misconception Tailored dialogue 180
Figure 10.5. Learner’s characteristic curve 181
Figure 10.6. Measuring emotional responses during tutoring 184
Figure 11.1. Issues affecting measurement equivalence across cultural and linguistic groups 193
Figure 12.1. Three key uses of process data 212
Figure 12.2. Multimodal data capture to reflect thinking processes 221
Figure 14.1. Task sequence in the PISA 2025 Learning in the Digital World assessment 249

TABLES

Table 1.1. Challenges for the assessment of 21st Century competencies 35
Table 4.1. Problem-solving decisions in science and engineering 83
Table 4.2. Comparison of scientific literacy in PISA and problem-solving decisions in S&E 86
Table 5.1. How technology can innovate assessment 101
Table 6.1. Design patterns for the practice of modelling in the PISA 2025 LDW assessment 122
Table 7.1. Sources of evidence in technology-enhanced assessment 141
Table 8.1. A taxonomy of conceptual elements at the intersection of measurement science and learning analytics 150
Table 8.2. Construct-relevant information available in original version of PILA Karel task for accumulation on the ‘nested loop’ goal 155
Table 9.1. Affordances of digital learning resources and opportunities to elicit and evaluate SRL 161
Table 9.2. SRL inferences and supporting evidence 164
Table 10.1. Response types and input devices 176
Table 10.2. Indices of group communication derived from interactions in an ITS generic scenario 182

BOXES

Box 2.1. Examples of invention activities using contrasting cases as learning resources 52
Box 2.2. Catering to different student ability groups in PILA 54
Box 3.1. Simulating the role of a city planner in SimCityEdu 65
Box 3.2. Betty's Brain: Searching for and representing information in an open learning environment 68
Box 3.3. Assessing students' design work in E-Scape 71
Box 3.4. Collaborative tasks in large-scale assessments: examples from PISA and ATC21S 72
Box 8.1. Applying hybrid models to the Harvard VPA task "There's a New Frog in Town" 153
Box 11.1. Optimising comparability in the PISA 2022 Creative Thinking assessment 194

Editorial

More than 20 years on from its first cycle, PISA 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 learned 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 21st 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 now provides us 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 OECD Secretary-General

Executive summary

One of the most important ways that assessments can function positively in education systems is by signalling the competencies that matter and illustrating the types of performances we want students to master. Because what we choose to assess inevitably ends up being taught in classrooms, assessing what matters – and doing it well – should be a priority for education policy. This volume makes the case for pursuing innovation in assessment in terms of the types of educational outcomes we should assess, the way we design tasks – capitalising on technology to generate rich and meaningful sources of data – and the processes required to ensure that assessments are valid given their intended use.

Assessments should measure what matters, not just what is easy to measure

What is worth knowing, doing and being has been subject to an intense debate over recent decades. Educational stakeholders agree that we need to support the development of complex cognitive and socio-cognitive constructs (or "21st century competencies"). Although frameworks describing these skills share similarities, translating this vision into practice requires aligning curriculum, pedagogy and assessment. Assessment can drive this alignment but challenges include defining constructs and learning progressions, developing tasks that elicit valid evidence and defining suitable models to interpret and report evidence.

Next-generation assessments should enable students to demonstrate what they can do in authentic contexts and evaluate how students learn new things

Better assessing 21st century competencies requires "next-generation assessments". Insights from the learning sciences suggest several assessment design innovations that align with this goal including: using extended performance tasks with "low floors, high ceilings"; situating assessments in authentic contexts of practice; including opportunities for exploration, discovery and invention; and including feedback and learning scaffolds. During assessment, students should be given opportunities to engage in the types of learning, decision making and problem solving engaged by practitioners in the real world.

Because 21st century competencies are strongly intertwined in practice, creating separate assessments for individual competencies may not be a productive strategy. Decisions about what to assess might be better guided by three interrelated considerations: 1) identifying a cluster of relevant activities that require students to engage in learning, problem solving and decision making; 2) identifying the context of practice and the disciplinary or cross-disciplinary knowledge required in that context; and 3) deciding whether to integrate affordances to enable students to work collaboratively or independently.

Innovation is necessary across all phases of assessment design

Assessment design is always an exercise in design science: tasks and interpretation methods must be anchored to a well-defined theoretical framework for assessments to generate valid inferences. This is

especially the case for next-generation assessments of complex skills. Valid inferences about students' capacity to engage in complex types of problem solving and learn new things should combine top-down (justified by theory) and bottom-up (visible in data) arguments and evidence. This requires close collaboration among potential users of assessment, domain experts, psychometricians, task designers, software designers and user interface (UI) experts from the outset.

Digital technologies vastly expand the assessment designer's toolbox, but new and better measurement models are needed

Digital technologies enable new and innovative task formats (including interactive and immersive problems and environments), test features (including adaptivity and affordances for learning) and potential sources of evidence (including work products or solutions, as well as a wide array of process data capturing student behaviours and processes). Although these new sources of now data are relatively "easy" to obtain from technology-enhanced assessments, existing psychometric models do not handle the complexity of these types of data well. New measurement models are needed, especially at scale – for example, exploring "hybrid" measurement solutions that incorporate one measurement model within another.

Intelligent Tutoring Systems (ITS) that provide students with dynamic tasks, interactivity and feedback might provide useful inspiration for how to develop choice-rich tasks and innovative scoring methods. Many ITS have made advances using artificial intelligence (AI)-based technologies, such as natural language processing, to provide intelligent feedback to learners, adapt content in response to their actions and evaluate what they know and can do. While learning analytics methods are increasingly intersecting with educational measurement, gaps remain to be bridged between both fields in order to use these new methods in a way that truly benefits the users of assessment.

Next-generation assessments require careful validation through principled design processes and data collection and evaluation

Complex constructs are inevitably shaped by cultural norms and expectations. In large-scale assessment, generating valid evidence through complex tasks must be balanced with the need to achieve score comparability. New issues that are specific to innovative digital assessments (e.g. the relationship between digital literacy and performance, potential biases in AI-based methodologies) must also be thoroughly evaluated. It is critical that evidence to support equivalence is established both through principled design processes and through dedicated empirical studies. Process data represent a valuable source of validity evidence concerning how individuals and different student groups engage with an assessment.

Next-generation assessments require intellectual, fiscal and political investment

Developing next-generation assessments will require the simultaneous investment of several types of capital: intellectual, involving different communities of experts in learning science, measurement science and data science to collaborate and solve conceptual and technical challenges; fiscal, to support the multidisciplinary teams required to design innovative assessments and bring promising examples to scale; and political, to buy into the vision to invest beyond the current possible and transform entrenched practices in assessment, and to assemble the fiscal capital required. International large-scale assessment programmes, like PISA, can play a pioneering role in bringing together these three capitals and innovating assessment at scale.

Introduction: Arguments in support of innovating assessments

By James W. Pellegrino

(University of Illinois Chicago)

This introduction establishes assessment as a process of reasoning from evidence and presents the main arguments for why we need to innovate assessments. The first argument is that assessment should measure what matters, not just what is easy to measure. This means expanding the range of educational outcomes we assess to include the complex cognitive and socio-cognitive constructs that students will need for the worlds of today and tomorrow. The second argument is that we need new assessment designs that leverage the affordances of digital technology to provide rich and meaningful sources of data. Following from the first two arguments, the third is that assessments should measure what matters and measure it well. Careful attention must be paid to the issues of validity and comparability when complex constructs are targeted for assessment, and when new tasks and tools are used for generating and interpreting evidence about student performance.

Overview

This introduction sets forth the main arguments for innovating assessments that are elaborated across the chapters in this report. The first argument is that educational policy and practice need to (re)consider what is important to measure and better define the various components of what are often complex constructs and the authentic contexts in which we engage them. In education we need to *measure what matters*, not simply what is *easy to measure*. The second argument follows from the first – to assess constructs that matter we need to innovate the ways in which we design assessments and the technologies we use to assist in this process – all while bearing in mind the goal of generating useful evidence about what students know and can do with respect to these constructs. The third argument follows from the first two – for the results of any such assessments to be useful to the intended audiences, be they teachers, administrators or policy makers, they must be valid (i.e. assess those competencies that they purport to measure and not others) and they must be comparable (i.e. assess those competencies reliably across assessment contexts and socio-cultural groups). Furthermore, the particular user group(s) need to be able to make sense of the results. Thus, reporting of the forms of evidence generated by innovative assessments must be done in ways that accurately reflect the complexity of the constructs being assessed and the intended uses of the information.

The three interconnected arguments noted above broadly motivate the division of this report into three distinct parts: 1) the 'what' of assessment; 2) the 'how' of assessment; 3) and the interpretation and use of results from innovative assessments, including considerations of reliability and comparability. To develop and elaborate these arguments we begin with a brief discussion of a fundamental conception about assessment, namely that it constitutes a process of reasoning from evidence guided by theory and research on critical aspects of knowledge and skill. This fundamental principle provides a basis for developing each of the three arguments noted above, including their elaboration in subsequent chapters of this report. We conclude this chapter with an additional argument of consequence for educational policy and practice – to achieve innovation in assessment and effect positive impact on educational outcomes, more coherent systems of assessment are needed. Such systems better connect assessments to one another given their intended interpretive uses and their relationship to curriculum and instruction, respectively.

Assessment as a process of reasoning from evidence

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

The process of collecting evidence to support inferences about what students know and can do represents a chain of reasoning from evidence about student competence that characterises all assessments, from classroom quizzes and standardised tests, to computerised tutoring programmes, to the conversation a student has with her teacher as they work through a math problem or discuss the meaning of a text. The first question in the assessment reasoning process is: "evidence about what?" *Data* become *evidence* in an analytic problem only when one has established their relevance to a conjecture being considered (Schum, 1987, p. 16[1]). Data do not provide their own meaning; their value as evidence can arise only through some interpretational framework. What a person perceives visually, for example, depends not only on the data she receives as photons of light striking her retinas but also on what she thinks she might see. In the present context, educational assessments provide data such as written essays, marks on answer

sheets, presentations of projects or students' explanations of their problem solutions. These data become evidence only with respect to conjectures about how students acquire knowledge and skill.

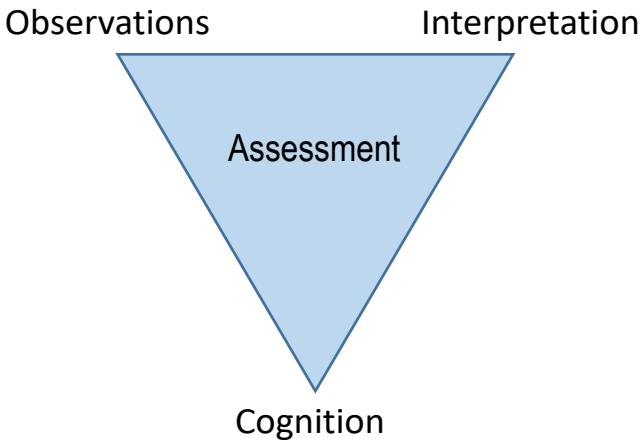
In the Knowing What Students Know report (Pellegrino, Chudowsky and Glaser, 2001[2]), the process of reasoning from evidence was portrayed as a triad of three interconnected elements: the *assessment triangle*. The vertices of the assessment triangle represent the three key elements underlying any assessment (see Figure A): 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 in light of the assessment purpose and student understanding. These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, or evaluated, without consideration of each. The three are represented as vertices of a triangle because each is connected to and dependent on the other two. A major tenet of the Knowing What Students Know report is that for an assessment to be effective and valid, the three elements must be in synchrony. The assessment triangle provides a useful framework for analysing the underpinnings of current assessments to determine how well they accomplish the goals we have in mind, as well as for designing future assessments and establishing their validity (Marion and Pellegrino, 2007[3]; Pellegrino, DiBello and Goldman, 2016[4]).

Figure A. 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, Chudowsky and Glaser (2001[2]).

The *cognition* corner of the triangle refers to theory, data and a set of assumptions about how students represent knowledge and develop competence in an intellectual domain (e.g. fractions, Newton's laws or thermodynamics). In any particular assessment application, 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 future learning. A central premise is that the cognitive theory should represent the most scientifically credible understanding of typical ways in which learners represent knowledge and develop expertise in a domain.

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

be carefully designed to provide evidence that is linked to the cognitive model of learning and to support the kinds of 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.

Every assessment is also based on certain assumptions and models for interpreting the evidence collected from observations. The *interpretation* vertex of the triangle encompasses all the methods and tools used to reason from fallible observations. It expresses how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed. In the context of large-scale assessment, the interpretation method is usually a statistical model, which is a characterisation or summarisation of patterns one would expect to see in the data given varying levels of student competency. In the context of classroom assessment, the interpretation is often made less formally by the teacher and is often based on an intuitive or qualitative model rather than a formal statistical one. Even informally, teachers make coordinated judgments about what aspects of students' understanding and learning are relevant, how a student has performed 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 not only must make sense on its own, but also must connect to each of the other two elements in a meaningful way to lead to an effective assessment and sound inferences. Thus, to have a valid and effective assessment, all three vertices of the triangle must work together in synchrony.

Argument 1: Measuring what matters

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

There is an increasing push to encourage students to develop "21st Century skills" that combine habits of mind and that include social and affective competencies (Bellanca, 2014[5]; Pellegrino and Hilton, 2012[6]). The European Commission's Rethinking Education (2012[7]) reform effort emphasises the need to promote transversal skills in education, such as critical thinking and problem solving. Additionally, the Programme for International Student Assessment (PISA) – the international assessment of student abilities administered by the OECD – has begun testing broader competencies that go beyond the disciplinary areas of mathematics, reading and science such as problem solving and collaborative problem solving. Such 21st Century skills – or 21st Century competencies, as referred to throughout this report – are deemed necessary to prepare a global workforce to succeed in a new information-driven economy. Individuals must have the problem solving, critical thinking, and collaboration and communication skills to evaluate and make sense of new information and to act upon this information in a range of settings.

Business leaders, educational organisations and researchers have begun to call for new education policies that target the development of such broad, transferable skills and knowledge. For example, the US-based Partnership for 21st Century Skills (2010[8]) argues that student success in college and careers requires four essential skills: critical thinking and problem solving, communication, collaboration, and creativity and

innovation. The report Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century (Pellegrino and Hilton, 2012[6]) argued that the various sets of terms associated with the "21st Century skills" label reflect important dimensions of human competence that have been valuable for many centuries, rather than skills that are suddenly new, unique and valuable today. The important difference across time may lie in society's desire for all students to attain levels of mastery – across multiple areas of skill and knowledge – that were previously unnecessary for individual success in education and the workplace. At the same time, the pervasive use of new digital technologies has increased the pace of communication and information exchange throughout society with the consequence that all individuals may need to be competent in processing multiple forms of information to accomplish tasks that may be distributed across contexts that include home, school, the workplace and social networks.

In order to shift from policy into practice, assessments need to be able to measure these skills and competencies. To do that we need to have clear conceptions and definitions of the constructs to be assessed (i.e. the *cognition*), the forms of evidence associated with those constructs (i.e. the *observations*), and ways to make sense of that evidence for the purposes of reporting and use (i.e. the *interpretation*).

This report's first four chapters explicitly focus on the 'what' of educational assessment – the key constructs that we should be interested in assessing, why those constructs are important, and where we stand with respect to assessing them given the current educational assessment landscape. The bulk of the argument across Chapters 1-4 is that we should be focused on complex cognitive and socio-cognitive constructs, both within and across disciplinary domains. The chapters discuss what we mean by these constructs and the types of tasks and situations where individuals would be required to exercise the requisite competencies, thereby providing the types of evidence that would be valid, interpretable and useful whether the intended use is at the classroom level to guide learning and instruction or in a large-scale educational monitoring context. Each of the chapters illuminate ways in which we might conceptualise and operationalise these constructs, as well as some of the challenges in doing so. They set the stage for chapters that follow on moving from conceptualisation of what we may want and need to assess as part of the advancement of 21st Century education, to the details of the design process and ways in which technology can enable the creation of situations that will provide the evidence we need while also assisting in the process of making sense of that evidence.

Argument 2: Assessment design processes and applications of technology

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

Recognising that assessment is an evidentiary reasoning process, it has proven useful to be more systematic in framing the process of assessment design as an Evidence-Centred Design (ECD) process (Mislevy and Haertel, 2007[9]; Mislevy and Riconscente, 2006[10]). The process starts by defining the claims that one wants to be able to make about student knowledge and the ways in which students are supposed to know and understand some particular aspect of a content domain. Examples might include aspects of algebraic thinking, ratio and proportion, force and motion, heat and temperature, etc. The most critical

aspect of defining the claims one wants to make for the purposes of assessment is to be as precise as possible about the elements that matter and express these in the form of verbs of cognition that are much more precise and less vague than high-level cognitive, superordinate verbs such as know and understand. Example verbs might include compare, describe, analyse, compute, elaborate, explain, predict, justify, etc. Guiding this process of specifying the claims is theory and research on the nature of domain-specific knowing and learning.

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

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

In the Knowing What Students Know report (Pellegrino, Chudowsky and Glaser, 2001[2]), many of the affordances of technology for advancing assessment design and practice were discussed in terms of the three interconnected components of the assessment triangle. The brief discussion that follows focuses on the constructs that could be represented in innovative assessment frameworks (cognition), the ways in which those constructs could be realised in the assessment environment (observations), and some of the interpretive challenges and solutions associated with doing so for purposes of measurement and reporting (interpretation).

The cognition vertex of the assessment triangle

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

The observation vertex of the assessment triangle

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

the intended reasoning and response processes that form the basis for desired claims about student proficiency (Gorin and Mislevy, 2013[11]).

Students' interactions with these technology-enhanced assessments can be logged to provide data on how they engage in particular processes. For various 21st Century competencies, the process by which one completes the activity can be as important a piece of information about knowledge and skill as the final product. In these cases, understanding the operations that students performed in the process of creating the final product may be critical to evaluating students' proficiency. Log data offer the opportunity to reveal these actions, including where and how students spend their time and what choices they make in situations like using a simulation. Such applications offer the potential to provide large volumes of "click-stream" and other forms of response process data that might be useful for making inferences about student thinking (Ercikan and Pellegrino, 2017[12]).

The interpretation vertex of the assessment triangle

Technology offers significant opportunities to enhance the reasoning-from-evidence process given the types of observations described above. Collecting these types of data makes little sense unless there are ways to reliably and meaningfully interpret them. This can evolve through mechanisms such as automated scoring of responses and application of complex parsing, statistical and inferential models for response process data (Ercikan and Pellegrino, 2017[12]). Critical data to consider include the time taken to perform various actions, the actual activities chosen, and their sequence and organisation. The potential exists for examining the global and local strategies students use while solving assessment problems and their implications, including how such strategies relate to the accuracy or appropriateness of final responses. Although capturing such data in a digital environment is relatively easy, making sense of the data is far more complicated. The same can be said for capturing data to constructed response questions where students may be expressing in written and/or graphical form an argument or explanation about some social, economic or scientific problem or phenomenon, describing the design of an investigation, or representing a model of some structure or process.

The data capture contexts described above are challenging regarding scoring and interpretation. It is here that artificial intelligence and machine learning may play a significant role in future innovative assessments (Zhai et al., 2020a[13]; 2020b[14]). Developments in machine learning also may allow researchers to analyse complex response process data, including to reveal patterns that provide important insights into students' cognitive processes in problem solving (Zhai et al., 2020a[13]; 2020b[14]; 2021[15]; Zhai, 2021[16]; Zhai, Krajcik and Pellegrino, 2021[17]). Such data may prove to be especially informative about student thinking and reasoning and thus add greatly to the knowledge gained about student competence from large-scale assessments like PISA. An interesting example was provided in a recent report by Pohl et al. (2021) who showed that differences in student response processes, when combined with scoring methods, can significantly change the interpretation of a country's performance in PISA.

In summary, digital technologies hold great promise for helping to bring about the changes in assessment that many believe are necessary. Technologies available today and innovations on the immediate horizon can be used to access information, create simulations and scenarios, allow students to engage in learning games and other activities, and enable collaboration among students. Such activities make it possible to observe, document and assess students' work as they are engaged in natural activities – perhaps reducing the need to separate formal, external assessments from learning in the moment (Behrens, DiCerbo and Foltz, 2019[18]). Technologies will certainly make possible the greater use of formative assessment that in turn has been shown to significantly impact student achievement. Digital activities may also provide information about abilities such as persistence, creativity and teamwork that current testing approaches cannot. Juxtaposed with this promise is the need for considerable work to be done on issues of scoring and interpretation of evidence before such embedded assessment can be useful for these varied purposes.

Developing assessments of complex cognitive competencies requires being explicit about all three elements of the assessment triangle and their inter-relationships. While Chapters 1-4 of this report primarily focus on Argument 1 concerns regarding the *cognition* element of the assessment triangle, Chapters 5-10 address various aspects of Argument 2 regarding the *observation* and *interpretation* elements of the assessment triangle, with an emphasis on how technology can be exploited through and within a principled design process to create assessments of the complex cognitive and socio-cognitive performances that matter. Through a combination of argument and specific examples, Chapters 5-10 provide support for the claim that next-generation assessments are possible but can only be generated through a highly principled design process that makes explicit the evidentiary chain of reasoning at the core of valid assessment. The chapters also reveal the complexities that accrue in designing such assessments and then making sense of the multiple forms of evidence they can produce.

Argument 3: Valid interpretation and use of results

The joint American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) Standards (1999[19]; 2014[20]) frame validity largely in terms of "the concept or characteristic that a test is designed to measure" (2014, p. 11[20]). In Messick's construct-centred view of validity, the theoretical construct the test score is purported to represent is the foundation for interpreting the validity of any given assessment (Messick, 1994[21]). For Messick, validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (1989, p. 13[22]). Important work has been done to refine and advance views of validity in educational measurement (Haertel and Lorie, 2004[23]; Kane, 1992[24]; 2001[25]; 2006[26]; 2013[27]; Mislevy, Steinberg and Almond, 2003[28]). Contemporary perspectives call for an interpretive validity argument that "specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances" (Kane, 2006, p. 23[26]).

Kane (2006[26]) and others (Haertel and Lorie, 2004[23]; Mislevy, Steinberg and Almond, 2003[28]) distinguish between: 1) the interpretive argument, i.e. the propositions that underpin test score interpretation; and 2) the evidence and arguments that provide the necessary warrants for the propositions or claims of the interpretive argument. In essence this view identifies as the two essential components of a validity argument the claims being made about the focus of an assessment and how the results can be used (interpretive argument), together with the evidence and arguments in support of those claims. Appropriating this approach, contemporary educational measurement theorists have framed test validity as a reasoned argument backed by evidence (Kane, 2006[26]). An argument and evidence framing of validity supports investigations for a broad scope of assessment designs and purposes, including many that go beyond typical large-scale tests of academic achievement or aptitude and move one into the arena of innovative and instructionally supportive assessments (Pellegrino, DiBello and Goldman, 2016[4]).

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

evidence. Such evidence can be derived from multiple forms of data involving variations in human performance and are essential to establishing an assessment's validity argument.

In pursuing innovative assessments of 21st Century competencies, of paramount concern are issues of equity and fairness as part of the validity argument. Of particular concern is the comparability of results and validity of inferences derived from performance obtained across different modes of assessment, especially for varying groups of students (Berman, Haertel and Pellegrino, 2020[29]). As large-scale assessment has moved from paper-and-pencil formats to digitally-based assessment, the general focus has been on mode comparability and concerns about student familiarity and differential access to the hardware and software used (Way and Strain-Seymour, 2021[30]). However, as the digital assessment world advances, a significant issue for large-scale innovative assessment is determining how student background characteristics including language, culture and educational experience influence performance on different types of tasks and innovative assessment designs that leverage the power of technology. As the assessment environments and tasks become more innovative, equity and fairness concerns become even more important than general mode comparability effects. Thus, a key part of the validity argument for any innovative assessment will be establishing the socio-cultural boundaries related to equitable and fair interpretations and uses of the assessment results.

Much of this report focuses on critical aspects of design and development as part of establishing the validity of next-generation assessments for 21st Century competencies. More specifically, Chapters 5-10 focus on the validity evidence that would be derived through the application of a principled design process that forces one to articulate, in varying degrees of detail, the connections between and among the *cognition*, *observation* and *interpretation* components of the assessment. Such evidence contributes to the assessment's overall validity argument but needs to be complemented by various forms of empirical data on how the assessment performs. Chapters 11-13 extend the validity evidence and argument discussion by considering comparability and fairness concerns in large-scale, technology-rich assessments, as well as considering the valid interpretation and use of results derived from innovative analytic approaches. These chapters discuss methodologies and principles for examining validity issues throughout assessment design and once assessment data have been collected.

Towards more coherent systems of assessment

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

At least four major features define the elements of assessment systems that can fully reflect rigorous standards and support the evaluation of deeper learning (see Darling-Hammond et al. (2013[31]) for an elaboration of the relevance, meaning and salient features of each of these criteria):

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

Tasks should measure these abilities directly as they will be used in the real world rather than through a remote proxy.

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

A major challenge is determining the conditions and resources needed to create coherent systems of assessments that work across contexts ranging from the classroom to larger organisational units such as districts, states, countries and internationally. Regardless of their context of implementation, assessments in such systems must support the ambitious goals we have for the educational system, meet the information needs of different stakeholders, and align with the criteria above. Aspects of this assessment system design and implementation challenge are taken up in the Conclusion chapter of this report.

Conclusion

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

This report is focused on an alternative and perhaps revolutionary vision that starts with the complex cognitive competencies that are deemed critical for citizens of the 21st Century. The report's chapters provide a vision of what they are by characterising how we might create environments and situations where the competencies of interest would necessarily be expressed in addition to describing the evidence that those environments could provide about those competencies. Some might find it curious that a vision for the future of education starts with assessment rather than curriculum and instruction. One of the benefits of thinking first about the outcomes we desire from the educational system, with a particular focus on what they would look like, is that this information provides the basis for a 'Backwards Design' process regarding the design of curriculum and instruction that can lead to those outcomes (Wiggins and McTighe, 2011[32]).

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

References

AERA, APA, NCME (2014), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, D.C., https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf. [20]

AERA, APA, NCME (1999), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, D.C. [19]

Behrens, J., K. DiCerbo and P. Foltz (2019), "Assessment of complex performances in digital environments", *The ANNALS of the American Academy of Political and Social Science*, Vol. 683/1, pp. 217-232, https://doi.org/10.1177/0002716219846850. [18]

Bellanca, J. (2014), *Deeper Learning: Beyond 21st Century Skills*, Solution Tree Press, Bloomington. [5]

Berman, A., E. Haertel and J. Pellegrino (eds.) (2020), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, National Academy of Education, Washington, D.C., https://doi.org/10.31094/2020/1. [29]

Darling-Hammond, L. et al. (2013), *Criteria for High-Quality Assessment*, Stanford Center for Opportunity Policy in Education, Stanford. [31]

Ercikan, K. and J. Pellegrino (eds.) (2017), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, https://doi.org/10.4324/9781315708591. [12]

European Commission (2012), *Rethinking Education: Investing in Skills for Better Socio-Economic Outcomes*, European Commission, Strasbourg. [7]

Gorin, J. and R. Mislevy (2013), "Inherent measurement challenges in the Next Generation Science Standards for both formative and summative assessment", *Paper presented at the Invitational Research Symposium on Science Assessment, Washington D.C.*, Washington, D.C., https://www.ets.org/Media/Research/pdf/gorin-mislevy.pdf. [11]

Haertel, E. and W. Lorie (2004), "Validating standards-based test score interpretations", *Measurement: Interdisciplinary Research & Perspective*, Vol. 2/2, pp. 61-103, https://doi.org/10.1207/s15366359mea0202_1. [23]

Kane, M. (2013), "Validating the interpretations and uses of test scores", *Journal of Educational Measurement*, Vol. 50/1, pp. 1-73, https://doi.org/10.1111/jedm.12000. [27]

Kane, M. (2006), "Validation", in Brennan, R. (ed.), *Educational Measurement*, American Council on Education/Praeger, Westport. [26]

Kane, M. (2001), "Current concerns in validity theory", *Journal of Educational Measurement*, Vol. 38/4, pp. 319-342, https://doi.org/10.1111/j.1745-3984.2001.tb01130.x. [25]

Kane, M. (1992), "An argument-based approach to validity", *Psychological Bulletin*, Vol. 112/3, https://www.act.org/content/dam/act/unsecured/documents/ACT_RR90-13.pdf. [24]

Marion, S. and J. Pellegrino (2007), "A validity framework for evaluating the technical quality of alternate assessments", *Educational Measurement: Issues and Practice*, Vol. 25/4, pp. 47-57, https://doi.org/10.1111/j.1745-3992.2006.00078.x. [3]

Messick, S. (1994), "The interplay of evidence and consequences in the validation of performance assessments", *Educational Researcher*, Vol. 23/2, pp. 13-23, https://doi.org/10.2307/1176219. [21]

Messick, S. (1989), "Meaning and values in test validation: The science and ethics of assessment", *Educational Researcher*, Vol. 18/2, pp. 5-11, https://doi.org/10.3102/0013189x018002005. [22]

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. [9]

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*, Lawrence Erlbaum, Mahwah. [10]

Mislevy, R., L. Steinberg and R. Almond (2003), "On the structure of educational assessments", *Measurement: Interdisciplinary Research and Perspectives*, Vol. 1/1, pp. 3-67. [28]

Partnership for 21st Century Skills (2010), *21st Century Readiness for Every Student: A Policymaker's Guide*, https://files.eric.ed.gov/fulltext/ED519425.pdf (accessed on 4 March 2023). [8]

Pellegrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academy Press, Washington, D.C., http://faculty.wiu.edu/JR-Olsen/wiu/common-core/precursor-documents/KnowingWhatStudentsKnow.pdf. [2]

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. [4]

Pellegrino, J. and M. Hilton (eds.) (2012), *Education for life and work: Developing transferable knowledge and skills in the 21st century*, The National Academies Press, Washington, D.C., https://doi.org/10.17226/13398. [6]

Schum, D. (1987), *Evidence and Inference for the Intelligence Analyst*, University Press of America, Lantham. [1]

Way, D. and E. Strain-Seymour (2021), "A framework for considering device and interface features that may affect student performance on the National Assessment of Educational Progress", *Paper commissioned by the NAEP Validity Studies Panel*, https://www.air.org/sites/default/files/Framework-for-Considering-Device-and-Interface-Features-NAEP-NVS-Panel-March-2021.pdf. [30]

Wiggins, G. and J. McTighe (2011), *The Understanding by Design Guide to Creating High-Quality Units*, ASCD. [32]

Zhai, X. (2021), "Practices and theories: How can machine learning assist in innovative assessment practices in science education", *Journal of Science Education and Technology*, Vol. 30/2, pp. 139-149, https://doi.org/10.1007/s10956-021-09901-8. [16]

Zhai, X. et al. (2020a), “From substitution to redefinition: A framework of machine learning-based science assessment”, *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1430-1459, https://doi.org/10.1002/tea.21658. [13]

Zhai, X. et al. (2021), “A framework of construct-irrelevant variance for contextualized constructed response assessment”, *Frontiers in Education*, Vol. 6, pp. 1-13, https://doi.org/10.3389/feduc.2021.751283. [15]

Zhai, X., J. Krajcik and J. Pellegrino (2021), “On the validity of machine learning-based Next Generation Science Assessments: A validity inferential network”, *Journal of Science Education and Technology*, Vol. 30/2, pp. 298-312, https://doi.org/10.1007/s10956-020-09879-9. [17]

Zhai, X. et al. (2020b), “Applying machine learning in science assessment: A systematic review”, *Studies in Science Education*, Vol. 56/1, pp. 111-151, https://doi.org/10.1080/03057267.2020.1735757. [14]

Part I Innovating What We Assess

1 21st Century competencies: Challenges in education and assessment

By Natalie Foster

(OECD)

This chapter reviews several frameworks of so-called "21st Century competencies", reviewing their main ideas and the vision of education they seek to promote. It also elaborates on the interrelations between 21st Century competencies – both among each other and to disciplinary learning – before discussing some of the key challenges that this vision of education presents for contemporary education systems. In particular, this chapter discusses challenges in the context of their assessment, including defining assessment constructs and learning progressions, generalisability of assessment claims, task and item design, interpreting scoring and evidence, reporting and validation.

Introduction

What is worth knowing, doing and being has been subject to a global conversation since before the turn of the 21st Century. Success in global contemporary society demands a wider set of competencies that go beyond the traditional literacies of reading, mathematics and science. Information and communication technologies (ICTs) have radically transformed our societies connecting people around the world and delivering unprecedented amounts of information to us, in turn giving rise to new forms of decentralised and autonomous learning. Young people today must not only learn to participate in a more interconnected, digital and rapidly changing world, they must also learn to develop their own agency and make decisions that contribute to individual and collective well-being. They need to understand and appreciate different perspectives, interact and collaborate successfully with others, and take responsible action towards creating a cohesive and sustainable future for all.

The nature of work is also changing. A declining proportion of the labour market in OECD economies is engaged in jobs consisting of routine work and manual labour (OECD, 2016[1]). The past two decades have seen a significant shift towards economies and societies that increasingly rely on human knowledge to produce new goods and services, underpinned by skills such as creative thinking, innovation and complex problem solving.

These significant trends have important consequences for schooling, teaching and learning. The knowledge and skills that today's students need to thrive in rapidly changing labour markets and to live with others as responsible, democratically- and socially-engaged citizens are changing. Value has shifted away from memorising content towards developing interdisciplinary skills (so-called "21st Century skills" or competencies) and acquiring deeper learning outcomes or transferable knowledge.

Over the past 20 years 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 active and effective participation in the emerging global knowledge society (Pellegrino and Hilton, 2012[2]; Fadel and Groff, 2018[3]; Binkley et al., 2011[4]; Scott, 2015[5]; World Economic Forum, 2015[6]; European Commission, 2019[7]). The OECD's Learning Framework 2030 goes one step further, emphasising the need to cultivate students' agency as a key goal of a 21st Century education so that young people are able to fulfil their potential and actively contribute to the well-being of their communities and the planet (OECD, 2018[8]). A number of works have also comparatively analysed such frameworks (Voogt and Roblin, 2012[9]; Scott, 2015[5]; Chalkiadaki, 2018[10]; Joynes, Rossignoli and Amonoo-Kuofi, 2019[11]).

While the literature reveals a general consensus on what 21st Century competencies are and why they are important, there remain significant challenges to the adoption of this agenda in practice. First, there is no singularly agreed-upon approach for identifying the competencies that should be prioritised in formal education nor how specific competencies are defined or delimited in relation to others. Second, shifting a greater focus on the development of 21st Century competencies throughout formal education requires accompanying shifts across curricula, pedagogy and assessment, as well as ensuring that these systems closely align. Yet several open questions remain about how best to learn, teach and assess 21st Century competencies in the classroom.

This chapter begins by examining frameworks of 21st Century competencies. It discusses the twin issues of a lack of theory and definition and the need for education system alignment in some detail, with a particular focus on their implications for educational assessment. It then identifies six interconnected assessment challenges for developing assessments of 21st Century competencies: 1) construct and learning progression definition; 2) generalisability; 3) task and item design; 4) interpreting and scoring evidence; 5) reporting; and 6) validation. The chapter concludes by summarising the implications of these interconnected assessment challenges in terms of the requirements they impose on next-generation assessments of 21st Century competencies.

What are 21st Century competencies?

Before going further, it is useful to establish what exactly the term "21st Century competencies" means. 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 also extends to the ways in which different frameworks identify and define specific competencies (e.g. ICT literacy vs. digital literacy vs. media literacy). For the sake of clarity, this chapter uses the term 21st Century competencies to refer to the broad vision of education set forth by the frameworks cited above and to the various competencies that they describe. They are generally understood to refer to the knowledge, skills and attitudes necessary to be successful for living and working in the 21st Century global knowledge economy, to participate appropriately in an increasingly diverse society, to use new technologies effectively, and to adapt to change and uncertainty.

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[9]).

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[2]). Others group competencies according to their purpose or context of use, for example ways of thinking, ways of living in the world, ways of working and tools for working (Binkley et al., 2011[4]). Abstracting from the specificities of each framework, some broadly distinct categories of competencies do consistently emerge (Figure 1.1). While these six broad categories capture the essence and exhaustive lists of competencies identified across different frameworks, note that not all frameworks include each category, nor do they always assign specific competencies to the same broader categories.

Figure 1.1. Broad categories of 21st Century competencies

Identifying common categories of competencies provides some useful insight into the broader goals of education that these frameworks seek to promote, but they nonetheless remain strongly interlinked in the sense that engaging one "type" of competence often requires engaging other "types" simultaneously. For example, problem solving (usually categorised as a cognitive competence) also requires individuals to

monitor their progress and adapt accordingly (i.e. metacognitive competencies) and likely some degree of persistence (i.e. intrapersonal competence) in order to reach a successful solution.

Regardless of the way in which specific competencies are categorised across the 9 frameworks reviewed here, critical thinking, creative thinking, communication and ICT-related competencies are consistently identified as those that young people need to develop. This also largely reflects the 21st Century competencies that are most commonly cited within national curricula documentation (Care, Anderson and Kim, 2016[12]). All frameworks identify the importance of civics and citizenship, although some regard these as a cross-cutting knowledge area (along with others like financial, health, environmental and global literacies) rather than a specific type of competence. Most frameworks also identify problem solving, collaboration, metacognition and self-regulated learning, as well as some intrapersonal competencies.

The intersection of digital competencies with other 21st Century competencies is also addressed within each framework. This is because the proliferation of ICTs in both personal and professional life forms one of the central arguments for developing 21st Century competencies, primarily through their functionalities that enhance our capacity for communication, collaboration, use of knowledge and information-finding. However, frameworks differ in how they conceptualise digital competencies: some consider them to be their own distinct category of competencies (Partnership for 21st Century Learning, 2019[13]; Binkley et al., 2011[4]), whereas others employ more integrative approaches where the development and use of ICT knowledge and skills is embedded within other 21st Century competencies such as critical thinking, problem solving, communication or collaboration (Voogt and Roblin, 2012[9]).

A multi-faceted challenge for education systems

Nearly one quarter of the way through the 21st Century, the idea of "21st Century competencies" is no longer particularly new. So why has there not been more progress towards achieving this vision of education? In their analysis of national education documents in 102 countries, Care, Anderson and Kim (2016[12]) found that although a majority of countries acknowledge 21st Century competencies within their broader educational vision statements there is wide variation across countries in terms of the specific skills or competencies they reference. One major obstacle is the lack of a clear and universal definition of 21st Century competencies that moves beyond general rhetoric – one that specifies both what is included within this umbrella term and how those competencies relate (or not) to one another.

There is also clearly a lack of shared understanding of how these competencies develop and how they can be taught. Even if most countries acknowledge 21st Century competencies in some way, relatively fewer explicitly integrate them within their curricula or provide clear developmental progressions for them (Care, Anderson and Kim, 2016[12]). This in turn provides little practical guidance to educators in terms of designing and implementing educational approaches to teaching, learning and assessing these competencies. A second major challenge is therefore how to actually implement a 21st Century competencies agenda in practice. Real change towards adopting this agenda has implications across all aspects of education systems – from curriculum to pedagogy to assessment – and requires all three to be well-aligned.

Challenges in curriculum and pedagogy

While a more in-depth discussion about curriculum and pedagogy is beyond the scope of this chapter, they are nonetheless central components of any 21st Century competencies agenda and are intricately connected to how these competencies can and should be assessed. One ongoing debate in the field concerns disciplinary versus interdisciplinary approaches to teaching and learning 21st Century competencies. While these competencies are widely accepted as being interdisciplinary, what it means to problem solve, think critically or be creative in one context may be very different in another context. In other

words, being able to successfully engage 21st Century competencies when embedded in different domains depends, at least to some extent, on a foundation of relevant knowledge in that domain. Pellegrino and Hilton (2012, p. 4[2]) suggest that 21st Century competencies represent "transferrable knowledge" that comprises both "content knowledge in a domain and also procedural knowledge of how, why, and when to apply this knowledge".

In a review of the literature on four different 21st Century competencies, Lai and Viering (2012[14]) concluded that domain-specific knowledge is an important requisite – although to different extents depending on the specific competence. Critical thinking, for example, requires a foundation of domain-specific knowledge (with some even refuting the existence of domain-general critical thinking processes), while experts tend to agree that creativity has both domain-specific and domain-general components (Lai and Viering, 2012[14]). Yet even for competencies with some domain-general components, the context in which they are engaged can influence the degree to which those components are relevant. For example, both convergent and divergent thinking processes are important for creative thinking across domains, but convergent thinking might be relatively more important in scientific or engineering domains than in artistic domains.

This has implications for both curriculum and pedagogy. First, curriculum designers need to know what foundational knowledge facilitates the development and application of 21st Century competencies within different domains. Curricula need to integrate 21st Century competencies in the context of particular content knowledge and to treat both as equally important. Second, and related, more effort needs to be directed towards defining developmental progressions so that students are taught the right foundational knowledge to support the integration of 21st Century competencies at the right time – and so teachers can be informed about what to reasonably expect from students at different stages of education. When students first encounter new ideas or concepts, their understanding is shallow and often bound to specific examples. Students need to develop and organise their conceptual knowledge and understanding sufficiently to facilitate its application to novel situations by engaging 21st Century competencies. The teaching strategies that allow students to do this integrate carefully-designed direct instruction with hands-on inquiries that actively engage students in using material they have learnt with higher-order competencies of increasing complexity (Darling-Hammond et al., 2019[15]). In an iterative process, learning that engages higher-order competencies allows knowledge to be understood deeply enough to be recalled and used for other purposes in novel situations (Learning Policy Institute and Turnaround for Children, 2021[16]).

All of this means that opportunities to engage 21st Century competencies must be integrated systemically and strategically throughout the curriculum. One way to do this is through student-centred learning methods, such as problem-based or project-based learning, that empower students to work on real-world problems in authentic and genuinely meaningful ways. These participatory approaches to learning enable students to research and evaluate information using different resources and to actively construct their own knowledge and skills instead of passively absorbing and memorising information. Real-world problems are also rarely confined to a single content area, which make them ideal contexts for engaging interdisciplinary 21st Century competencies. Problem-based and project-based learning approaches encourage students to make connections between content areas and engage competencies like critical and creative thinking, problem solving and collaboration (Paniagua and Istance, 2018[17]).

Despite the fact that student-centred pedagogical approaches are widely acknowledged as beneficial, they impose additional demands on educators. They require that teachers invest time in devising engaging, student-centred lesson plans connected to the curriculum and manage a more interactive classroom in which students collaborate with each other and engage in autonomous research. At the system level, this requires a greater investment in teacher development and training, as well as ensuring that educators are given the autonomy to integrate the curriculum in ways that make sense for their classrooms and that they feel empowered to do so.

Challenges in assessment

There is little point in investing heavily in curriculum and educator training reform without investing in assessment to evaluate what is (or is not) being accomplished in the classroom. Curricula, pedagogy and assessment are intricately connected and must be aligned in well-functioning education systems. Assessments – especially large-scale – are important signposts indicating what students should learn and what they can do. Shifts in curricula and pedagogy can thus be driven by changes in an education system's assessment focus and by the educational gaps that they reveal, in turn informing policymaking and reform. Moreover, explicitly focusing assessment on these competencies requires that they are clearly defined and requires being specific about what exactly they involve at different educational levels, therefore contributing to establishing a shared understanding of these competencies and how they should be taught.

However – despite some promising examples at scale (e.g. the Programme for International Student Assessment) – there is currently a lack of systemic understanding in how to measure or capture the attainment of 21st Century competencies (Care et al., 2018[18]; Vista, Kim and Care, 2018[19]). This is because several of the challenges related to defining or integrating 21st Century competencies in curriculum and pedagogy have similarly complex implications for assessment. Six major interconnected assessment challenges, described below in Table 1.1, are discussed in detail in the remainder of this chapter. These assessment challenges are: 1) defining constructs and learning progressions; 2) generalisability; 3) task and item design; 4) interpreting and scoring evidence; 5) reporting; and 6) validation.

Table 1.1. Challenges for the assessment of 21st Century competencies

Challenge	Source of complexity	Implication(s) for assessment
Defining constructs and learning progressions	21st Century competencies are complex and multidimensional, involving cognitive, metacognitive and affective processes. They are also strongly interconnected when engaged authentically. Robust learning progressions are also generally lacking.	• The constituent variables need to be clearly defined at the beginning of any assessment design process to inform task design, the evidence to be collected, and the claims to be made about student performance. • It can be difficult to isolate and interpret evidence for discrete constructs. • It can be difficult to identify processes and outcomes (i.e. evidence) linked to different levels of mastery.
Generalisability	21st Century competencies require some degree of domain-specific knowledge to be meaningfully engaged, and they may also be understood and defined differently in different domains.	• Domain-general assessments lack validity. • The greater the domain-specificity of an assessment, the weaker the generalisability of its claims about students' capacity to engage 21st Century competencies outside of that specific domain context.
Task and item design	Students need to work on more open-ended, interactive and authentic tasks in order to engage 21st Century competencies and demonstrate their proficiency.	• Simple item types cannot fully reflect the range of authentic scenarios that engage complex constructs nor capture the range of evidence of proficiency (e.g. interpersonal competencies are evidenced primarily through behaviours when interacting with others). • Multiple instruments or item types are generally required to gather information about all relevant aspects of the construct.
Interpreting and scoring evidence	Students need to work on more open-ended, interactive and authentic tasks in order to engage 21st Century competencies. Processes and behaviours are also key aspects of performance.	• Evidence generated by process data can be challenging to interpret and use for scoring. • Open-ended items without a pre-determined list of "correct" responses may require human scoring (with implications in terms of time and cost).
Reporting	21st Century competencies are complex and multidimensional, involving cognitive, metacognitive and affective processes.	• Unidimensional score scales are inadequate, but multiple scales (or scales with clearly distinct dimensions) are difficult to achieve given the constraints of large-scale assessments (e.g. limited testing time). • Some constructs might not be best described by linear point scales, but the exploration of alternative reporting methods has been scarce so far.
Validation	21st Century competencies are complex and multidimensional, involving cognitive, metacognitive and affective processes. Processes and behaviours are also often associated with performance.	• How constructs are defined and how student performance is interpreted are more susceptible to cultural bias than traditional tests of knowledge. • More open, interactive and process-oriented task and scoring models impose additional burdens in terms of establishing a validity argument.

Defining constructs and learning progressions

21st Century competencies are complex and multidimensional constructs. They involve a combination of cognitive, metacognitive and affective processes, and are supported by a set of knowledge, skills and attitudes. These multiple components are often strongly inter-related with other 21st Century competencies and most authentic, real-world contexts require individuals to engage several competencies simultaneously – making it difficult to clearly distinguish mastery in one competency from another. For example, problem solving involves aspects of metacognition, self-regulated learning and persistence – and depending on the context and typology of the problem, it could also involve elements of creative thinking and collaboration, as well as domain-specific knowledge (see *Generalisability* section below).

For any assessment to fully represent and measure its target construct, all of the constituent elements of the construct must be clearly defined and captured reliably through the assessment instrument(s). As set forth in the Introduction, a sound theoretical framework must describe the kinds of evidence that need to be generated and collected to sustain claims about students' performance and to design appropriate tasks and define proficiency scales that reflect the different levels of competence mastery (Wilson et al., 2011[20]; Ercikan and Oliveri, 2016[21]). However, the complexity of these constructs and the conceptual crowding within the broader discourse on 21st Century competencies means it is difficult to break down constructs into discrete and independently measurable components, as well as isolate and attribute evidence generated by students to one particular competence or another (Ercikan and Oliveri, 2016[21]).

This definition issue is challenging not only in terms of identifying exactly what to measure but also in terms of how to interpret performance. A lack of well-defined learning progressions mean it is harder to identify the processes and outcomes that students demonstrate at different levels of competence mastery, as well as to design and locate assessment tasks that sample students' competency at various levels of complexity or sophistication. This is particularly relevant in the context of assessing 21st Century competencies since significant evidence is also expressed through students' behaviours and processes not only their final output (Care et al., 2018[18]).

Generalisability

While 21st Century competencies can be applied across subject domains, they are all bound by some extent to the context in which they are applied. This tension between domain-specificity and domain-generality has important implications for assessment. First, assessment designers need to be clear about whether and how the target construct changes across domains – for example, whether one aspect of the construct is relatively more important in some domains of application compared to others – and they need to design items that can elicit relevant kinds of evidence accordingly. This of course relies on a strong foundation of theory about the nature of the construct both within and across domains.

Second, the role and importance of domain-specific knowledge in the assessment needs to be considered, with important trade-offs in terms of developing authentic tasks and making generalisable claims about student performance. One approach is to limit the relevance of domain knowledge to the target construct by situating tasks in neutral and accessible contexts or by focusing measurement on domain-general processes. However, this limits the authenticity of the tasks and assessment claims as 21st Century competencies are rarely exercised in reality in contexts where no relevant knowledge is beneficial. Providing the knowledge that test takers need directly within the task prompt could be one way to mediate this limitation but test takers with existing, well-organised knowledge schemas or knowledge of domain-relevant strategies might nonetheless have an advantage.

A different approach is to acknowledge that it is neither possible nor desirable to disentangle 21st Century competencies from domain-specific knowledge and instead integrate their measurement within domain-specific assessments (see Chapter 4 of this report for an example of this approach) – although this has consequences on the generalisability of inferences made about student performance beyond the given

assessment domain (Ercikan and Oliveri, 2016[21]). One way to mediate the limitations of this approach is to employ a sampling design: in other words, recognise that knowledge is a relevant component of 21st Century constructs and develop assessment tasks across several domain contexts so that the assessment can provide a more comprehensive view of students' strengths and weaknesses across domains (Lai and Viering, 2012[14]). This approach clearly has practical implications in terms of developing sufficient items across contexts and gathering enough information from students, both within and across different domains, to be able to draw valid and reliable conclusions about student performance.

Item characteristics and task design

Most large-scale assessments rely on traditional item formats, such as multiple choice, true/false statements or close-ended responses, to elicit indicators about students' underlying abilities. While static and close-ended items are easy to code and score, they inherently limit what can be captured about students' performance and essentially target the reproduction of content knowledge. For constructs like mathematics knowledge, the link between test indicators and construct is fairly direct: a correct response demonstrates knowledge of the topic. But these items are not optimal for generating indicators that capture the complexity and multi-component nature of 21st Century competencies, especially as these constructs are defined (at least in part) by behaviours or processes (Lai and Viering, 2012[14]; Care et al., 2018[18]; Vista, Kim and Care, 2018[19]).

Assessments of 21st Century competencies need to generate evidence that indicates not just what students know but also how they deal with complex situations and iterate towards a solution. For example, "good" self-regulated learning or collaboration is characterised as much by behaviours, attitudes and ways of thinking as it is by eventual (successful) outcomes. The challenge for assessment is that simple indicators of knowledge or outcomes do not capture well these underlying processes. As such, test items for measuring 21st Century competencies need to focus on making students' behaviours and thought processes visible (Lai and Viering, 2012[14]).

This key issue requires innovation in task and item design in two interconnected ways. First, assessment tasks must mirror the kinds of authentic situations that require 21st Century competencies – not only to stimulate the processes and behaviours from which to generate indicators of the construct (Care, Anderson and Kim, 2016[12]) but also to ensure that claims about student abilities actually reflect performance in real-life contexts (Care et al., 2018[18]; Lai and Viering, 2012[14]; Ercikan and Oliveri, 2016[21]). The literature generally agrees that 21st Century competencies empower individuals to address new and complex problems and adapt to unpredictable situations, meaning authentic tasks are best situated within the context of open-ended, ill-structured problems (see also Chapter 2 of this report). What's more, some types of 21st Century competencies require situations that involve others with whom to interact (e.g. collaboration, communication) or that trigger some kind of personal state, emotion or level of investment (e.g. persistence, conflict resolution, self-regulated learning). These types of authentic problem situations are clearly more difficult to generate within the constraints of a controlled test environment and in a way that engages all students to the same extent (i.e. that elicits evidence from all students).

Second, and closely related, test items need to be open-ended and interactive so that they can make visible test takers' behaviours and thinking processes. For many 21st Century competencies, this means providing students with tools for doing and making and a test environment that enables them to engage in the entire process of idea conception to implementation by providing them with choices and opportunities to explore and iterate upon their ideas. These kinds of affordances cannot be sufficiently provided by the static, closed-response item types typically used in large-scale assessment. While technology provides new opportunities to address these needs (see Chapters 5 and 7 of this report), designing and validating technology-enhanced items also demands more time and financial resources.

Interpreting and scoring evidence

The scoring of items in any assessment is closely related to the task and item design, the claims that the assessment aims to make, and the definition of the construct and its learning progressions (Csapó et al., 2011[22]). Assessment tasks must elicit relevant evidence from students, but this evidence needs to be interpreted, clearly connected to the construct and levels of performance (as defined by construct maps or learning progressions) and accumulated using some kind of statistical model. Traditional assessments focusing on knowledge reproduction are easy to code and score: if students select or write the correct response then they receive credit. However, this simple scoring model can rarely be applied in the context of measuring 21st Century competencies that are largely defined by thought processes and behaviours, and for which it is not possible to define a concise, finite list of correct responses.

In recent decades, advances in computer-based assessment mean that evidence of student behaviours and thought processes can be captured through process (or log-file) data. But the interpretation of this evidence is far from straightforward: similar patterns of behaviour may mask real differences in thinking processes and approaches. For example, a prolonged period of recorded inactivity could be an indicator of disengagement or of a student who is deep in thought. Such behaviours may also be more susceptible to cultural differences, emphasising the importance of validation activities in computer-based assessments that make use of process data (see *Validation* section below, and also Chapter 12 of this report). Even when the interpretation of the behaviour is clear, it can be challenging to establish a scoring hierarchy among different behaviours and strategies as the optimal choice may depend on a host of other factors including the test taker's level of prior knowledge, motivation and even personality traits. For example, Roll et al. (2014[23]) demonstrated that while productive help-seeking was an optimal self-regulated learning strategy for most learners in an online problem-solving environment, trial-and-error was actually the most beneficial strategy for those with the lowest levels of prior knowledge.

New and more advanced analytical models are required to use evidence derived from process data, possibly in combination with more traditional types of outcome data. Authentic assessments of 21st Century competencies require test takers to engage in extended, performance-based tasks using interactive tools and resources – but these affordances introduce complexity in scoring and data analysis by giving test takers choice (meaning they may not experience the test in the same way, thus threatening comparability) and by introducing dependency across items (whereby students' prior decisions and actions determine later possibilities and outcomes). While making good choices during the assessment is a part of the construct, it may nonetheless confound other metrics. Innovative analytical models therefore need to account for these dependencies across items and the implications of additional constructs (such as choice) that are in play (see Chapter 8 of this report). New methods are also required to model changes in a students' proficiency over an extended task, as interactive and resource-rich test environments afford possibilities for students to learn over the course of the assessment.

It is not just process data that can pose interpretation and scoring challenges – for some 21st Century competencies it is also challenging to interpret students' outcome data. For example, in an assessment of creative thinking, there may be an infinite number of possible creative outcomes or solutions that cannot be pre-defined. In these cases, automated scoring methods are not possible without integrating some sort of sophisticated artificial intelligence or machine learning model; and human scoring of such complex and open-ended responses can also be unreliable and require a large investment in time and financial resources. Moreover, without clearly defined learning progressions for these competencies, it can be difficult to classify and score outcomes that reflect different levels of mastery.

Reporting

Once evidence from a test has been identified, collected and interpreted, it needs to be accumulated and reported in ways that are appropriate and useful for the intended purpose of the assessment. The reporting

of student performance in large-scale assessments is especially important as these assessments are used to inform policymaking and system-level reform (Vista, Kim and Care, 2018[19]).

21st Century competencies are complex and multidimensional constructs meaning that assessments intending to measure them should aim to make claims on students' ability across those different dimensions. One challenge in terms of reporting relates to how to generate evidence in such a way that the data scale together to provide an overall picture of student performance on the test, while at the same time providing insights on students' relative strengths and weaknesses. In large-scale assessment, test developers have prioritised achieving a single, reliable scale that summarises students' overall performance. Even when sub-scales are developed to measure different dimensions of the construct, these sub-scales are generally highly correlated with each other so that high (or low) performing students on the overall scale tend also to be high (or low) performing across all the sub-scales. More actionable information on strengths and weaknesses would require developing a more diverse set of item types that target the different dimensions of the construct, but this poses practical challenges including higher development costs and requiring a longer assessment time.

Another reporting challenge concerns how to describe complex behaviours in a way that provides actionable information to the users of the assessment data. For example, digital tests make it possible to record the different strategies that students adopt to solve complex, open problems. Analysing this process data can reveal different "profiles" of problem solvers, for example those who rapidly test ideas versus those who pause and reflect before attempting a solution. These data on solution processes provide a window into how students reason and construct their knowledge, giving potentially useful insights on the quality of instruction they have received. However, data on processes might be hard to convert into a linear point scale as some processes are not necessarily preferable to others. While some methodologies, such as cluster analysis, can be used to identify and describe profiles (e.g. different types of problem solvers), such reports are more complex to understand for policymakers and other users of the assessment data.

Validation

Innovative assessments of complex constructs like 21st Century competencies require more comprehensive validation processes that go beyond psychometric considerations of the reliability of scores (Vista, Kim and Care, 2018[19]; Care et al., 2018[18]; Ercikan and Oliveri, 2016[21]; Ercikan et al., 2016[24]). Like all assessments, innovative assessments need to be built on strong validity arguments that demonstrate that they measure what they intend to (construct validity), that student performance in the test is related to performance in relevant and authentic real-life situations (external validity), and – in large-scale assessment – that performance across student groups is comparable (fairness and cross-cultural and cross-linguistic comparability).

One set of challenges relates to whether 21st Century competencies are understood and expressed in similar ways across cultures and student groups. While it may be reasonable to assume that some competencies, like problem solving or critical thinking, involve similar thinking processes across cultures and student groups, others – such as those with stronger inter- or intra-personal components – are likely to be more sensitive to cultural or gender differences in terms of how they are expressed. This idea also extends to the ways in which different cultures or student groups may value different 21st Century competencies and consider them appropriate in certain situations (e.g. defining whether an output is creative or whether a problem requires a creative solution). It is therefore important that the target populations of any assessment of 21st Century competencies share a common understanding of the target construct to ensure there is a good degree of construct equivalence (Ercikan and Oliveri, 2016[21]).

Another set of challenges relates to validating test tasks and items and ensuring fairness among student groups. Many of these validation issues apply to all assessments: task contexts should aim to be equally familiar to students from different backgrounds, while avoiding socio-cultural, gender or other types of biases; task instructions (and translations, where applicable) should be expressed in the most appropriate

way to ensure that students clearly understand what is required of them; and response modes should be simple and intuitive so that construct-irrelevant factors do not unduly influence performance. In technology-enhanced assessments that provide more open and interactive test environments, the user interface design and user experience need careful attention so that the test remains accessible to all students. Students from different socio-cultural backgrounds may lack familiarity with such digital assessment platforms and their affordances, which may create sources of incomparability and jeopardise cross-cultural comparability. Cognitive laboratories and log data analysis can provide insight into students' thinking processes and test taking experiences that can be used for the purpose of validating tasks (see Chapter 12 of this report), but these exercises require a greater human and financial investment during the test development process (Ercikan and Oliveri, 2016[21]; Care, Anderson and Kim, 2016[12]).

The third set of validity challenges relates to the evidence and scoring models employed in innovative assessments of 21st Century competencies. A central tenet of score comparability across cultural and language groups is measurement invariance, which refers to the degree to which similar constructs are being measured and scores are comparable for test taker groups. In innovative assessments that integrate evidence from process data into scoring models, one challenge is ensuring that those processes are equivalent across student groups. However, students from different socio-cultural backgrounds or with different levels of prior knowledge may not use the tools and affordances of the test environment similarly, or they may not use the same strategies to solve tasks. This differential engagement can threaten the validity of conclusions and comparisons about student performance if evidence of those processes is interpreted and scored as evidence of mastery of the construct. International assessments further compound this challenge as linguistic or cultural differences may also lead to behavioural differences. For a comprehensive discussion of these issues in large-scale, innovative assessments of 21st Century competencies, see Chapters 11 and 12 of this report.

Finally, as technology-enhanced assessments are able to capture more complex process data, more complex and automated analyses for interpreting and scoring data using machine learning and artificial intelligence algorithms have also been developed (DiCerbo, 2020[25]). However, if the data sources used to create these scoring algorithms are not representative of all cultural and student groups then the resulting scores may not have equivalent validity and accuracy for all groups. One challenge for implementing such methods at scale is ensuring that all relevant student groups are adequately represented in training the algorithms, and that the scoring algorithms are sufficiently evaluated for validity, accuracy and comparability of scores for all of the diverse student populations.

Conclusion

21st Century competencies are increasingly recognised as key competencies for today's young people to develop so that they can effectively participate in the global knowledge economy, thrive in an increasingly diverse society, use new technologies effectively, adapt to change and uncertainty, and continue to engage in lifelong learning. Nonetheless, there remain significant challenges to the adoption of this agenda in practice across all aspects of the education system, from curriculum to pedagogy to assessment. This chapter focused on discussing six major challenges associated with designing valid assessments of complex 21st Century competencies. Doing so requires innovation throughout the entire assessment development process: from defining the conceptual framework, to task design, test delivery, validation, scoring, analysis and reporting.

Developing assessments of 21st Century competencies first requires clearly defining the target construct and establishing the theoretical underpinning of the assessment. The extent to which domain-specific knowledge supports the target construct also needs to be addressed in the conceptual framework as this will affect how assessment tasks are contextualised and the extent to which student performance in the assessment can be generalised. Without a clear conceptual framework, well-defined learning progressions

or construct maps, it is difficult to interpret and score student performance. This challenge is further amplified by the fact that 21st Century competencies are characterised as much by process as by outcome, that there is often no single or pre-defined “correct” response and that constructs are multidimensional (so performance may not be uniform across all dimensions). Because of the relative emphasis on students’ processes, assessment tasks need to be more open and interactive to provide opportunities for students to demonstrate how they engage in those processes. Task designers must therefore identify the types of complex yet accessible problems and situations that call for students to engage 21st Century competencies as well as develop test environments that allow students to respond in authentic ways and that generate interpretable evidence about their ways of thinking and doing. Finally, these challenges all pose additional demands in terms of establishing the validity argument for innovative assessments to ensure that tasks and scoring methods are equally accessible for different student groups and free from cultural, gender and linguistic bias.

References

Binkley, M. et al. (2011), "Defining twenty-first century skills", in Griffin, P., B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-007-2324-5_2. [4]

Care, E., K. Anderson and H. Kim (2016), *Visualizing the Breadth of Skills Movement Across Education Systems*, The Brookings Institute, Washington, D.C., https://www.brookings.edu/wp-content/uploads/2016/09/global_20160916_breadth_of_skills_movement.pdf (accessed on 23 February 2023). [12]

Care, E. et al. (2018), *Education System Alignment for 21st Century Skills: Focus on Assessment*, The Brookings Institute, Washington, D.C., https://www.brookings.edu/wp-content/uploads/2018/11/Education-system-alignment-for-21st-century-skills-012819.pdf (accessed on 12 February 2023). [18]

Chalkiadaki, A. (2018), "A systematic literature review of 21st century skills and competencies in primary education", *International Journal of Instruction*, Vol. 11/3, pp. 1-16, https://doi.org/10.12973/iji.2018.1131a. [10]

Csapó, B. et al. (2011), "Technological issues for computer-based assessment", in Griffin, P., B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-007-2324-5_4. [22]

Darling-Hammond, L. et al. (2019), "Implications for educational practice of the science of learning and development", *Applied Developmental Science*, Vol. 24/2, pp. 97-140, https://doi.org/10.1080/10888691.2018.1537791. [15]

DiCerbo, K. (2020), "Assessment for learning with diverse learners in a digital world", *Educational Measurement: Issues and Practice*, Vol. 39/3, pp. 90-93, https://doi.org/10.1111/emip.12374. [25]

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. [21]

Ercikan, K. et al. (2016), "Use of evidence-centered design in assessment of history learning", in Braun, H. (ed.), *Meeting the Challenges to Measurement in an Era of Accountability*, Routledge, New York, https://doi.org/10.4324/9780203781302-18. [24]

European Commission (2019), *Key Competences for Lifelong Learning*, Publications Office of the European Union, Luxembourg, https://data.europa.eu/doi/10.2766/569540 (accessed on 28 February 2023). [7]

Fadel, C. and J. Groff (2018), "Four-dimensional education for sustainable societies", in Cook, J. (ed.), *Sustainability, Human Well-Being, and the Future of Education*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-78580-6_8. [3]

Joynes, C., S. Rossignoli and E. Amonoo-Kuofi (2019), *21st Century Skills: Evidence of Issues in Definition, Demand and Delivery for Development Contexts*, Institute for Development Studies, Brighton, https://assets.publishing.service.gov.uk/media/5d71187ce5274a097c07b985/21st_century.pdf (accessed on 16 March 2023). [11]

Lai, E. and M. Viering (2012), "Assessing 21st century skills: Integrating research findings", *Paper presented at the National Council on Measurement in Education, Vancouver, B.C.*, Pearson, http://images.pearsonassessments.com/images/tmrs/Assessing_21st_Century_Skills_NCME.pdf. [14]

Learning Policy Institute and Turnaround for Children (2021), *Design Principles for Schools: Putting the Science of Learning and Development Into Action*, https://k12.designprinciples.org/sites/default/files/SoLD_Design_Principles_REPORT.pdf (accessed on 16 March 2023). [16]

OECD (2018), *The Future of Education and Skills 2030*, OECD Publishing, Paris, https://www.oecd.org/education/2030-project/about/documents/E2030%20Position%20Paper%20(05.04.2018).pdf (accessed on 1 March 2023). [8]

OECD (2016), "Automation and independent work in a digital economy", *Policy Brief on the Future of Work*, OECD Publishing, Paris. [1]

Paniagua, A. and D. Istance (2018), *Teachers as Designers of Learning Environments: The Importance of Innovative Pedagogies*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264085374-en. [17]

Partnership for 21st Century Learning (2019), *A Framework for Twenty-First Century Learning*, http://www.p21.org/ (accessed on 1 March 2023). [13]

Pellegrino, J. and M. Hilton (2012), *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, National Academies Press, Washington, D.C., https://doi.org/10.17226/13398. [2]

Roll, I. et al. (2014), "On the benefits of seeking (and avoiding) help in online problem-solving environments", *Journal of the Learning Sciences*, Vol. 23/4, pp. 537-560, https://doi.org/10.1080/10508406.2014.883977. [23]

Scott, C. (2015), "The futures of learning 2: What kind of learning for the 21st century?", *Education, Research and Foresight: Working Papers*, UNESCO, https://unesdoc.unesco.org/ark:/48223/pf0000242996. [5]

Vista, A., H. Kim and E. Care (2018), *Use of Data From 21st Century Skills Assessments: Issues and Key Principles*, The Brookings Institute, Washington, D.C., https://www.brookings.edu/wp-content/uploads/2018/10/EffectiveUse-Vista-Kim-Care-10-2018-FINALforwebsite.pdf. [19]

Voogt, J. and N. Roblin (2012), "A comparative analysis of international frameworks for 21st 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. [9]

Wilson, M. et al. (2011), “Perspectives on methodological issues”, in Griffin, P., B. McGaw and E. Care (eds.), *Assessment and Teaching of 21st Century Skills*, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-007-2324-5_3. [20]

World Economic Forum (2015), *New Vision for Education: Unlocking the Potential of Technology*, https://www3.weforum.org/docs/WEFUSA_NewVisionforEducation_Report2015.pdf (accessed on 16 March 2023). [6]

2 Next-generation assessments of 21st Century competencies: Insights from the learning sciences

By Mario Piacentini and Natalie Foster

(OECD)

Cesar A. A. Nunes

(Universidade Estadual de Campinas, São Paulo)

This chapter presents key insights from research in the learning sciences and from experimental evidence on instructional design to inform the design of next-generation assessments. It argues that research from the learning sciences emphasises that learning is a social process, and that expertise requires organised knowledge and the capacity to adapt and transfer that knowledge to novel situations. The chapter then sets forth several assessment design innovations that are aligned with these research insights, including using extended performance tasks that integrate opportunities for learning and iteration, that account for the role of knowledge in performance, and that have "high floors, low ceilings" to cater to different proficiency levels. It argues that such innovations may provide more valid information on how well educational experiences are preparing students for their future.

Introduction

In order to respond to the challenges of a world that is changing at an unprecedented speed, researchers, educators, policy makers and business leaders have emphasised the need to support 21st Century competencies. As discussed in the Introduction and Chapter 1 of this report, 21st Century competencies refer to the capacity to perform real-life activities such as analysing information critically, creating innovative solutions to problems, working productively with others and communicating effectively to different audiences. These competencies allow individuals to adapt effectively to changes in the labour market and society, and perform tasks on the job or roles within their communities that computers cannot replace (Levy and Murnane, 2004[1]).

The learning experiences that support these higher-order thinking and performance skills involve inquiry and investigation, the application of knowledge to new situations and problems, and the collaborative creation of ideas and solutions (Pellegrino and Hilton, 2012[2]). In these situations, students are invited to reflect on what is important to consider, plan what to do next and decide who to call on for help or feedback. The outcomes of these active learning experiences are often referred to as 'deeper learning', a type of learning whose positive effects transfer beyond the initial context of instruction (National Research Council, 2001[3]). 21st Century competencies constitute the integrated body of cognitive, intrapersonal and interpersonal skills that enable students to learn in ways that support not only information retention but also competent behaviour in other situations. As such, deeper learning experiences and 21st Century competencies are connected to each other in a self-reinforcing cycle: 21st Century competencies are developed in the context of deeper learning experiences, and deeper learning does not occur unless students can mobilise multiple 21st Century competencies simultaneously.

As argued in earlier chapters of this report, assessments are important markers of teaching and learning: if we want more students to engage in deeper learning experiences and to develop 21st Century competencies, then we need to develop new assessments that are capable of eliciting and measuring these competencies. Assessments at scale and across multiple cultural contexts can also help us to better understand these new constructs and how they interact in authentic problem situations. Assessments are more likely to yield valid evidence about what students know and can do if they confront students with the types of learning situations that actually require 21st Century competencies in real life, and such an alignment can generate evidence that differentiates between those students who have benefited from experiences of deeper learning at school from those who have not.

This chapter presents insights from research in the learning sciences on how expertise develops and from experimental evidence on instructional design that has implications for how to assess 21st Century competencies. The chapter then describes a number of design innovations for the next generation of assessments of 21st Century competencies that are aligned with these research insights and that have the potential to provide more valid information on how well educational experiences have prepared students for their future.

Research insights on deeper learning

It is increasingly recognised that expertise in a given domain is not limited to knowing facts and procedures, but extends to being able to organise this knowledge in a mental schema, to develop new solutions when established procedures fail or are unsuitable, and to communicate one's own emerging understanding and ideas to achieve a particular outcome (National Research Council, 2001[3]). Any assessment involves observing how an individual performs in only a handful of particular situations. Traditional large-scale assessments evolved to foster time and cost efficiency and compliance with established measurement models, resulting in a narrow focus on situations where test takers demonstrate the knowledge they already master rather than interacting with peers or mentors and building new knowledge. The focus of traditional

measurement approaches is also limited to the evaluation of response correctness – in other words, abstracting from the cognitive, interpersonal and intrapersonal processes that supported the response. As a result, the interpretation of a person's performance and behaviours in the types of restricted and static situations used in traditional assessments might not provide particularly valid inferences on that person's capabilities to think, act and learn in real-life situations outside of the test context.

It is possible to reduce this misalignment between assessments and contemporary theories of how people learn and solve problems in real life. This requires, as a first step, taking stock of the main findings from the research that has investigated the mental structures that support problem solving and learning. Some of this research has involved studies that contrast "experts" with "novices". In what follows, we use the word "expert" in relative terms to refer to individuals who have constructed solid mental models that connect ideas central to a given domain and who have learned, through participation in key practices, how to apply these ideas to new problems. We therefore do not necessarily refer to "experts" as established professionals with extensive experience in a specific discipline. Conversely, we refer to "novices" as those individuals who have learned the basics of a given domain but who have not had opportunities to consolidate their knowledge through practice.

The second step towards reducing this misalignment – and a much more difficult one – consists of creating an internally consistent system of teaching practices and assessments that reflect these research insights: in this system, deeper learning experiences prepare students for future learning and assessments measure how effectively students have engaged with these deeper learning experiences.

Research insight 1: Developing expertise is a social process

Over the last two decades, one of the main conclusions from research in the learning sciences is that learning is a socially situated process (Dumont, Istance and Benavides, 2010[4]; Darling-Hammond et al., 2019[5]). People develop expertise in a variety of domains, for example literature, engineering or the culinary arts, by participating in the practice of a community of experts and by learning to use the tools, languages and strategies that have been developed within that community (Mislevy, 2018[6]; Ericsson, 2006[7]; Pellegrino and Hilton, 2012[2]). The key to becoming skilful in a domain is acquiring fluency with the language and other cultural representations of the community, learning the ways of thinking and acting that are aligned with those representations, and soliciting and using feedback from other members of the community or from the situations themselves. Throughout life people learn interactively, trying different actions and observing their effects and consequences in the social space (Piaget, 2001[8]).

The fact that learning is mediated by socially-constructed practices has clear implications for instruction. Deeper learning occurs when students are given the opportunity to engage in activities that are realistic, complex, meaningful and motivating, and when they feel part of a community of learners that they can call on for support. In a responsive social setting, learners observe the criteria that others use to judge competence and adapt to these criteria. Innovative pedagogies have engaged students in highly interactive educational activities in which everyone is responsible for each others' learning. Some evidence shows that such practices, if carefully implemented by teachers, result in better academic results and higher enjoyment of learning – see, for example, the case of Railside school (Boaler and Staples, 2008[9]). The rewards and meaning that students derive from becoming deeply involved in collaborative knowledge building provide them with a strong motivation to learn and positive beliefs about what they can accomplish (Tan et al., 2021[10]).

Research insight 2: Experts engage in reflective practices and can adapt to new situations

A recurrent observation from studies that have compared experts to novices is that experts have strong metacognitive skills (Hatano, 1990[11]). In the course of learning and problem solving, experts display

regulatory behaviours such as knowing when to apply a procedure, predicting the outcomes of an action, planning ahead, monitoring their progress and efficiently apportioning cognitive and emotional resources. They also question limitations in their existing knowledge and avoid simple interpretations. This capability for self-regulation and self-instruction extends to situations when learning and problem solving happens in collaboration with others (Hadwin, Järvelä and Miller, 2017[12]).

The capacity to regulate one own's learning and (re)act accordingly is what distinguishes routine experts from adaptive experts, with the latter being "characterized by their flexible, innovative, and creative competencies within the domain" (Hatano and Oura, 2003, p. 28[13]). Adaptive experts are able to detect anomalies in their tasks and are consequently alerted when following established rules might result in sub-optimal outcomes; they also do not mind making errors in some situations, as errors teach them what not to do in particular situations and result in more integrated knowledge (Hatano and Inagaki, 1986[14]).

Research reveals that metacognition does not necessarily develop organically through traditional educational practices but that it can and should be explicitly taught in context (National Research Council, 2001[3]; Roll et al., 2007[15]). Reflecting on one's own learning – a major component of metacognition – does not typically occur in the classroom: when students are unable to make progress in their learning and are asked to identify the source of difficulty, they tend to report being "stuck" without analysing what they need to make progress. However, there are some notable examples of metacognitive strategies being used to improve learning in various domains. In mathematics, for example, teachers have had success with techniques that combine problem-solving instruction with control strategies for generating alternative problem-solving approaches, evaluating among several courses of action and assessing progress (Schoenfeld, 1985[16]).

Research insight 3: Expertise requires specialised and organised knowledge

Cognitive research has shown that general problem-solving procedures (or "weak methods") such as trial-and-error or hill climbing, are slow and inefficient (National Research Council, 2001[3]). Experts instead use deep knowledge of the domain ("strong methods") to solve problems. This deeper knowledge cannot be reduced to sets of isolated facts or propositions; rather, it is knowledge that has been encoded in a way that closely links it with its contexts and conditions of use. When experts face a new problem, they can readily activate and retrieve the subset of their knowledge that is relevant to the task at hand (Simon, 1980[17]). For example, chess experts encode mid-game situations in terms of meaningful clusters of pieces (Chase and Simon, 1973[18]). Research shows that students progress in their mastery of a discipline through similar processes of acquisition and use of increasingly well-structured knowledge schema.

The central role of domain knowledge for performance and future learning has strong implications for the teaching of 21st Century competencies. It would be unfair to expect that students can apply these competencies to a problem they know nothing about and cannot connect to their previous learning experiences. Finding a creative solution or communicating effectively about a topic generally require a deep understanding of relevant concepts and linguistic and socio-cultural patterns in the given domain. The teaching of these higher-order thinking and behavioural skills thus needs to be embedded within the conventions and "ways of knowing" of each learning area in the curriculum, and should possibly encourage students to establish connections between different disciplines.

Design innovations for new assessments of 21st Century competencies

As Chapter 1 of this report explained, we have a good understanding of what skills students need in order to learn and function in their future roles within society – but we do not yet have enough nor sufficiently adequate instruments to measure them. The objective of the following section is to highlight some general characteristics of what we consider to be "next-generation assessments" (NGA, for short), informed by the

above insights from the learning sciences, that can yield potentially valid evidence on where students are in their development of 21st Century competencies (in summative applications) and on what they need to do to progress in these skills (in formative applications).

Design innovation 1: Allow for extended performance tasks

Assessments that aim to measure how prepared students are for deeper learning have to engage students in active and authentic learning processes. As discussed earlier (see *Research insight 1*), students engage 21st Century competencies in situations in which they interact with others, evaluate available resources, make choices about what to focus on and disregard as well as the course of action to take, try out multiple strategies or iterations, and adapt according to the results. From an assessment perspective, this means providing students with a purposeful challenge that replicates the key features of those educational experiences where deeper learning happens as a result of interactions with the problem situation and knowledgeable others – including the capacity to make decisions (see Chapter 4 of this report for more on the importance of decision making in defining and assessing 21st Century competencies like problem solving).

In the context of summative assessments, in particular, efficiency considerations have led to short, discrete tasks being preferred over longer performance activities. In general, using many short items provides more reliable data on whether students master a given set of knowledge and can execute given procedures because 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 students' capacity to construct new knowledge in choice-rich environments then students should be given the time they need to demonstrate what they can do in these environments. This includes giving students time and affordances for reflective activities (*Research insight 2*).

Extended units that include multiple activities sequenced as steps towards achieving a main learning goal can provide students with a more authentic and motivating experience of assessment. 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 accomplish" – might ultimately provide more valid data (i.e. evidence that is predictive of what students are capable of doing outside of the constrained and stressful context of a test). These extended experiences are more challenging to design because developers need to establish a coherent storyline that keeps students engaged as well as address dependency problems (for example, by providing rescue points to move struggling students from one step to the next).

The OECD PISA 2025 test of Learning in the Digital World incorporates these ideas by including extended test units of 30 minutes. Each unit comprises various phases, including a learning phase (where students learn and apply concepts to simple problems) and a challenge phase (where students apply what they have learned earlier in the unit to solve a more complex problem). Another relevant model is the Cognitively-Based Assessment of, for and as Learning (CBAL) that uses scenario-based tasks designed by modelling high-quality teaching practices (Sabatini, O'Reilly and Wang, 2018[19]). A typical CBAL includes: 1) a realistic purpose; 2) sequences of tasks that follow learning progressions in a domain; and 3) learning material derived from multiple sources that reflect key practices in the content area.

Advances in technology (further explored in Part II of this report) allow much more data to be captured on how students spend their time in extended tasks by immersing them in simulated environments and communities of practice. These environments can facilitate a more open interpretation of students' goals and their exploration of the problem constraints, reward diverse solution strategies and outcomes, and provide feedback to learners. This also makes it possible to observe metacognitive processes that are crucial to learning in a non-obtrusive way, tracking how students plan and implement strategies, how they behave when they are stuck and how they respond to feedback (Nunes, Nunes and Davis, 2003[20]). The application of a principled design process can lead to a productive use of these process data to augment

the evidence that is derived from final solutions, therefore reducing the trade-off between reliability and authenticity (see Chapters 5-7 of this report).

Despite these advances, efficiency considerations (i.e. to collect as many observations as possible in limited testing time) will remain an important constraint. The argument here is thus not to shift completely from one assessment paradigm (i.e. using only short discrete items) to another (i.e. using only extended performance tasks) but rather encouraging the use of a more diversified set of assessment experiences where the breadth and depths of tasks and associated measurement models are aligned with what the assessment intends to measure.

Design innovation 2: Explicitly account for domain knowledge in assessment design and reporting

The socially-constructed nature of learning (*Research insight 1*) and the important role of knowledge for real-life performance (*Research insight 3*) both imply that we can hardly assess students' competencies like creativity, critical thinking or communication in a domain-neutral way. These skills are neither exercised nor observed within a vacuum. In an assessment context, students' ability to perform these skills will always be observed in a given 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 (Mislevy, 2018[6]). Attempting to design completely decontextualised assessment problems or scenarios threatens validity: if a student does not require any knowledge to solve a task, can the assessment truly claim to measure the types of complex competencies it claims to be interested in?

When designing assessments of 21st Century competencies, it is important to explicitly identify the knowledge students need to meaningfully engage with test activities and evaluate the extent to which differences in prior knowledge will influence the evidence we can obtain on the target skills. Moving beyond the idea of decontextualised assessments of 21st Century competencies can also help to make more valid claims when reporting what students can or cannot do. In the context of large-scale summative assessments in particular, it might be misleading to make general claims such as "students in country A are better problem solvers than students in country B". From a single summative assessment we might only be able to claim that "students in country A are better than students in country B at solving problems in the situations presented in the test" – most likely, a limited number of situations contextualised in one or few knowledge domains (see Chapter 3 of this report for a further discussion on how to define suitable assessment contexts for 21st Century competencies).

We suggest that measuring the relevant knowledge that students have when engaging with a performance task (for example through a short battery of items) should become an integral part of the design and assessment process in NGAs. This information can also help to interpret student's behaviours and choices in complex performance tasks (see Chapters 6 and 9 for a detailed presentation of the PISA 2025 Learning in the Digital World assessment that adopts this approach and considers students' actions in the context of what they already know).

Design innovation 3: Provide opportunities for productive failure and learning on the test

Metacognition, self-regulated learning and the capacity to flexibly apply one's knowledge to address new problems differentiates routine from adaptive experts (*Research insight 2*). Design choices related to the content and organisation of task sequences in NGAs can draw upon insights from research on instructional design for guidance on how to elicit these competencies in an assessment context. There is evidence that we can make robust claims about students' preparedness to learn new things by studying how they work on problems they have not encountered before (Roll et al., 2011[21]; Schwartz and Martin, 2004[22]).

One promising method involves "invention activities" that ask students to solve problems requiring concepts or procedures that they have not yet been taught, with the aim of encouraging students to invent

methods that capture deep properties of a construct before being taught expert solutions (Roll et al., 2012[23]). While inventing their own original approaches to solving novel problems, students tend to make mistakes and fail to generate canonical solutions. However, experimental evidence shows that students who learn through invention activities are better at transferring their knowledge (i.e. solving other tasks requiring the same knowledge schemes but in a different application) in comparison to students who are directly told what to do and then practice those procedures (Loibl, Roll and Rummel, 2016[24]; Kapur and Bielaczyc, 2012[25]). Invention activities therefore help students to deeply understand concepts, identify the limitations of previous interpretations and procedures when they do not work, and look for new patterns and interpretations that build upon and connect with their earlier knowledge.

Similar ideas can also be found in frameworks defining effective classroom practices, such as the "teaching for understanding" framework by Wiske (1997[26]). In the instructional context, these early attempts to solve problems prepare students to learn from subsequent instruction; in this way, they augment rather than replace instruction. When transposed to an assessment context these types of activities can provide important evidence about whether students can flexibly apply their knowledge schema to unfamiliar contexts as adaptive experts do.

Learning activities have to be carefully designed in order to support students in building their understanding while inventing and interacting with problems. In traditional tests, students are left to their own devices – they typically cannot draw upon anything other than their existing knowledge – and if they do not know the relevant procedure to follow there is little they can do to progress (Schwartz and Arena, 2013[27]). In the real world, we can access resources when learning and problem solving: we compare challenges to previous assignments, search the Internet for similar problems or solution strategies, or ask a knowledgeable other for directions (*Research insight 1*). Similarly, assessments that challenge students to create knowledge or solutions that are new to them should incorporate relevant resources for learning because problem solving always requires some degree of knowledge (*Research insight 3*). These resources should be carefully crafted so that they do not give the solution away but rather provide opportunities to learn more about the problem and the likelihood that trying a given solution or implementing a certain strategy will help them to make progress toward a solution.

Contrasting cases represent one approach to providing such structure in learning and invention activities that has proven effective in experimental settings and that could be applied to larger-scale assessments (see Box 2.1). Contrasting cases highlight key features that are relevant to specific decisions or concepts that students might overlook when addressing a problem, especially when they have nothing to compare to their current problem. With contrasting cases, students are encouraged to invent a representation that is general (i.e. works across different cases) rather than particular – thus encouraging knowledge transfer.

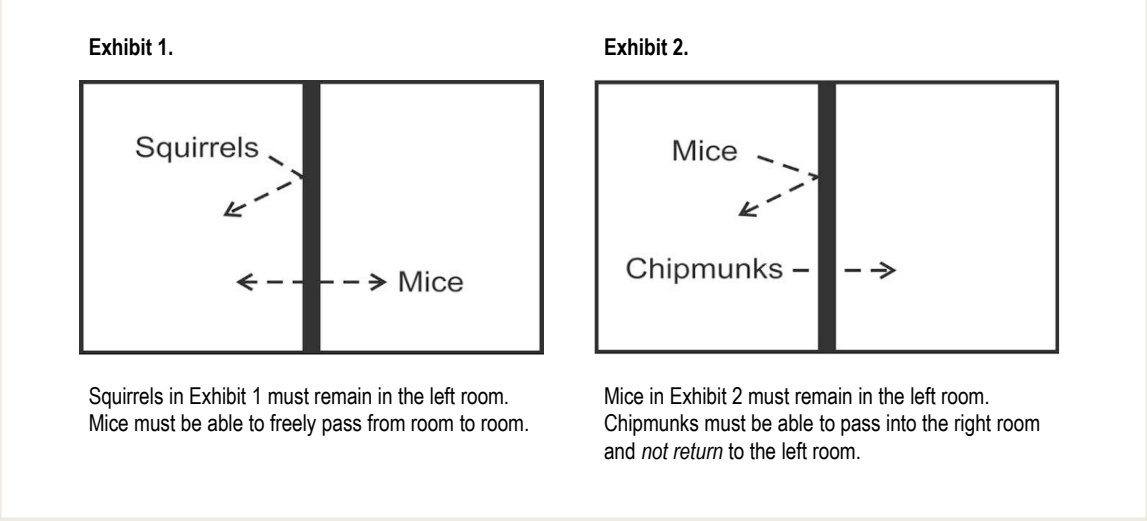
Integrating more than a single resource to learn from in an assessment task can provide additional evidence about whether, what and how students choose to learn. For example, Schwartz and Arena (2013) designed a game where young students are provided with a familiar problem (mixing colours) applied to an unfamiliar context (projecting lights onto a stage). Students had to learn that the result of mixing colours for light is different to mixing colours for paint, as light depends on a different set of primary colours (additive colours) than paint (subtractive colours). They had two tools from which to learn this concept – an experiment room, where they could directly mix light colours, and a set of catalogues (i.e. contrasting cases) – which made it possible to track how students decided to learn.

Box 2.1. Examples of invention activities using contrasting cases as learning resources

Schwartz and Martin (2004[22]) designed an invention activity for learning about applications of the mean deviation formula. Students had to find a procedure to identify which of four baseball-pitching machines was more reliable at aiming a target. Students were given four grids with different features including the number of balls shot by each machine. By comparing grids students could notice the effect of sample size on their measure, which helped them to understand why the standard deviation formula includes division by the number of observations.

In another experiment, first-year biology students were asked to design walls for a zoo exhibit of small rodents (Taylor et al., 2010[28]). Students were shown images of a squirrel, a chipmunk and a mouse, along with the approximate mass and dimensions of each rodent, and then provided with diagrams of two exhibits (see Figure 2.1). In both cases, the wall design needed to allow one rodent to freely pass from side to side while preventing the other rodent from doing so.

Figure 2.1. Contrasting cases example for designing a zoo exhibit



Squirrels in Exhibit 1 must remain in the left room. Mice must be able to freely pass from room to room.

Mice in Exhibit 2 must remain in the left room. Chipmunks must be able to pass into the right room and *not return* to the left room.

Source: Taylor et al. (2010[28]).

Students realised quickly that the wall in Exhibit 1 could use holes big enough for mice but too small for squirrels, but this simple solution did not work for Exhibit 2. Solution ideas for Exhibit 2 varied but typically students would use differences in mass, length of the chipmunks (to reach higher) or jumping ability to allow the chipmunks to cross. Some students also invented more original ideas, such as a tail scanner.

This problem is analogous to a problem biology students encounter in the study of living cells (e.g. the need for transport proteins to selectively allow certain molecules to cross a cell membrane) although on the surface it appears to have little to do with course material. The researchers found that students who engaged in these invention activities were much quicker to engage with unfamiliar biology problems and provided multiple reasonable hypotheses to explain unfamiliar problems compared to other students.

Design innovation 4: Provide feedback and instructional support during assessment

To complement providing resources for students to make sense of a problem and start inventing solutions (*Design innovation 3*), NGAs should also consider including guidance during the solution process in the form of advice, feedback or prompts. This type of instructional support can promote deep learning in beginners and enable the observation of decisions they make in their learning (Azevedo and Aleven, 2013[29]; van Joolingen et al., 2005[30]). Targeted feedback and interventions can also reduce the risk that beginners disengage from an assessment because they perceive it to be beyond their capacities – which is especially important in the context of more extended performance tasks (*Design innovation 1*).

Instructional support and feedback can play a variety of important functions including: 1) engaging a student's interest when they appear disengaged; 2) increasing their understanding of the requirements of a task when they demonstrate confusion; 3) reducing the degrees of freedom or the number of constituent acts required to reach a solution; 4) maintaining a student's direction; 5) signalling critical features including discrepancies between what a student has produced and what they recognise as correct; 6) demonstrating or modelling solutions, for example reproducing and/or completing a partial solution attempted by the student; and 7) eliciting articulation and reflection behaviour (Guzdial, Rick and Kehoe, 2001[31]).

In the context of large-scale assessments, feedback to students needs to be automated. This is clearly challenging because effective feedback is both task- and tutee-dependent: the feedback system must be based on a complete model of the task's demands, affordances and solution space, while also adapting to the performance of the student. Without attending to both, the system cannot generate feedback that is useful for all students, in turn failing to bring each student to their zone of proximal development (Vygotsky, 1978[32]). Artificial intelligence (AI) holds promise for answering to this challenge, at least for some types of learning and assessment experiences (see Chapter 10 of this report for more on AI-enabled adaptive feedback). Another challenge relates to the fact that students often do not proactively seek feedback or consult resources in interactive learning or assessment environments. It is therefore important to design such affordances in way that they are not too intrusive and distracting while being sufficiently visible and accompanied by explanations that invite test takers to use them.

Researchers have also invested considerable effort in creating multimedia learning environments that incorporate collaboration affordances as a way of supporting and monitoring knowledge-building processes (Lei and Chan, 2012[33]). In these environments, students receive feedback directly from their peers. Integrating these real-time, collaboration features in large-scale learning and assessment landscapes remains a challenge, although explorations are underway (see Rosé and Ferschke (2016[34]) as well as examples described in Chapter 3 of this report).

Including feedback and instructional support in summative tests is often challenged because of fairness concerns (e.g. concerns about unfairly penalising students who do not need support or threats to the validity of assessment claims). Shute, Hansen and Almond (2008[35]) undertook a rigorous evaluation of the psychometric quality of an assessment of algebra delivered by a digital learning system that combined adaptive task sequencing with instructional feedback. A comparison of metrics for a treated (with feedback) and control (without feedback) sample showed that providing instructional support did not make the assessment less able to detect differences between students or less valid. Indeed, providing feedback and support to students is more reflective of authentic learning experiences (*Research insight 1*). From a measurement perspective, the main challenge remains to develop and validate psychometric models for large-scale assessments that take into account how these resources affect students' measured ability, as this ability can no longer be considered as fixed but can progress as a result of using resources and feedback (see Chapters 6 and 8 of this report for a more in-depth discussion of this issue).

Design innovation 5: Cater to different ability levels by using challenges with "low floors, high ceilings"

In NGAs, all students should be able to demonstrate their ability to learn and progress by using the tools and resources available to them regardless of their initial level of knowledge or skill. Adapting assessment challenges to different abilities not only improves the quality of the measures but also the authenticity and attractiveness of the assessment experience. In real life people seldom take on challenges that they find either too easy or impossible to achieve, yet in traditional tests this happens quite frequently.

One approach to catering to different student ability levels involves designing tasks that have so-called "low floors, high ceilings", meaning that they are accessible to all students while still challenging top performers. Ensuring tasks are accessible for all students in terms of connecting to their previous learning experiences responds to the idea that successfully engaging 21st Century competencies requires some degree of relevant knowledge (*Research insight 3*). However, these types of problems are much more difficult to design than the standard problems found in traditional tests, where the item is matched to one specific level of difficulty and there is typically only one correct response.

One cluster of low floor, high ceiling problems asks students to produce an original artefact: this might 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 more complete, richer and unique. The low floor, high ceiling design can also be used in the context of more standardised problem solving tasks if students are told that there are intermediate targets to achieve and that they are expected to progress as much as they can towards a sophisticated solution (see Box 2.2 for an example assessment experience using a low floor, high ceiling approach to task design).

Box 2.2. Catering to different student ability groups in PILA

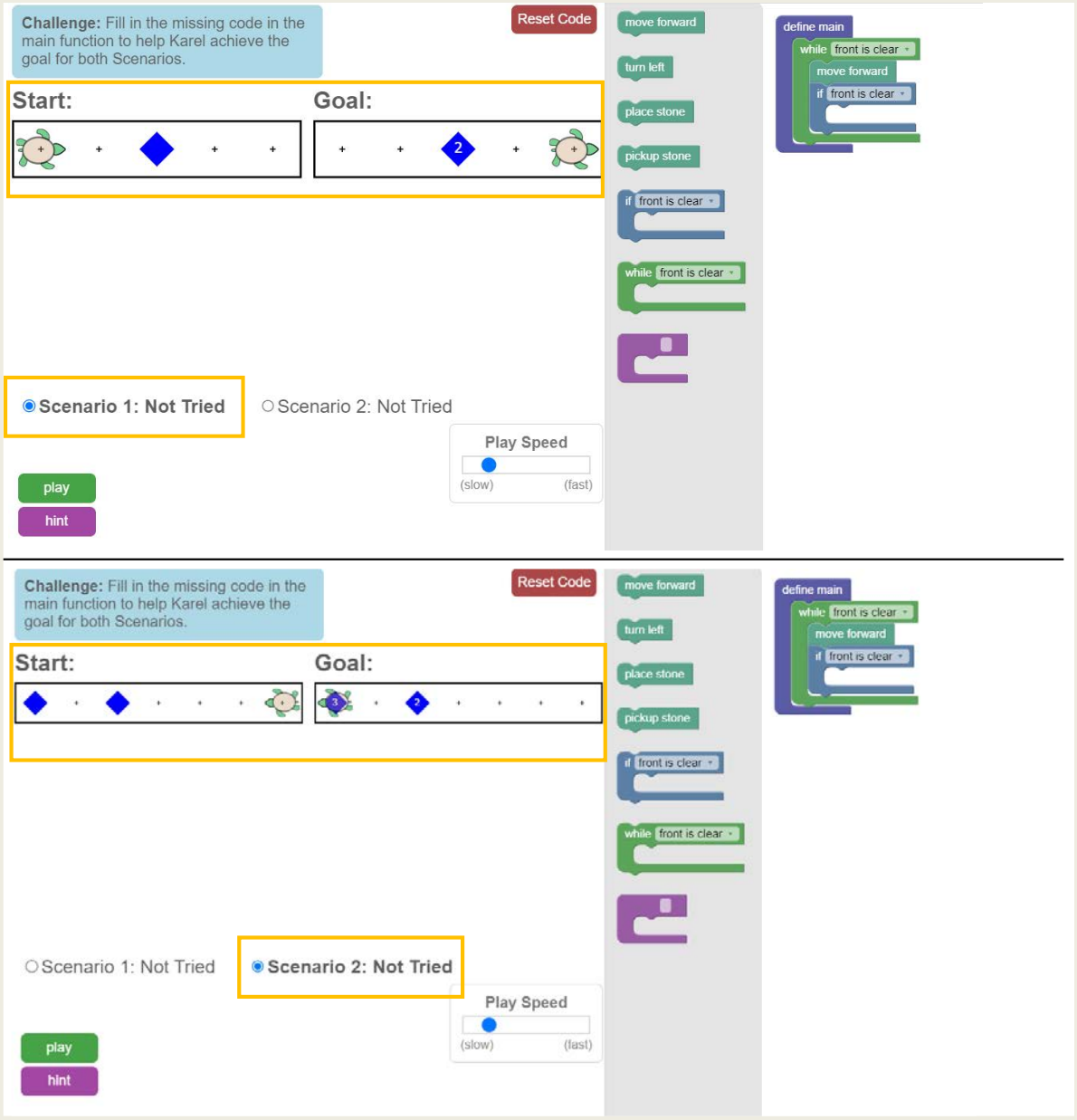
The Platform for Innovative Learning Assessments (PILA) is a research laboratory coordinated by the OECD. Assessments in PILA are designed as learning experiences that provide real-time feedback on student performance, typically for use in the context of classroom instruction. One overall objective of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

Designing low floor, high ceiling tasks for computational problem solving

One assessment 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. Tasks have a low floor, high ceiling: the intuitiveness of the visual programming language and the embedded instructional tools (e.g. interactive tutorial, worked examples) allow students with no programming experience to engage successfully with simple algorithmic tasks, yet the same environment can present complex problems that challenge expert programmers. Figure 2.2 shows an example task asking students to create a single program that moves Karel to the goal state in two different scenarios. To solve the problem, students can toggle between the two scenarios to visually observe the differences in the environment and how well their program works in both. Even students with solid programming skills generally require multiple iterations before finding an optimal solution for both scenarios, but the scoring models take into account partial solutions (e.g. solving the problem in only one scenario).

Figure 2.2. A low floor, high ceiling task in the PILA Karel application

"Two worlds" problem: the same program must solve the problem in both Scenario 1 (top) and Scenario 2 (bottom)



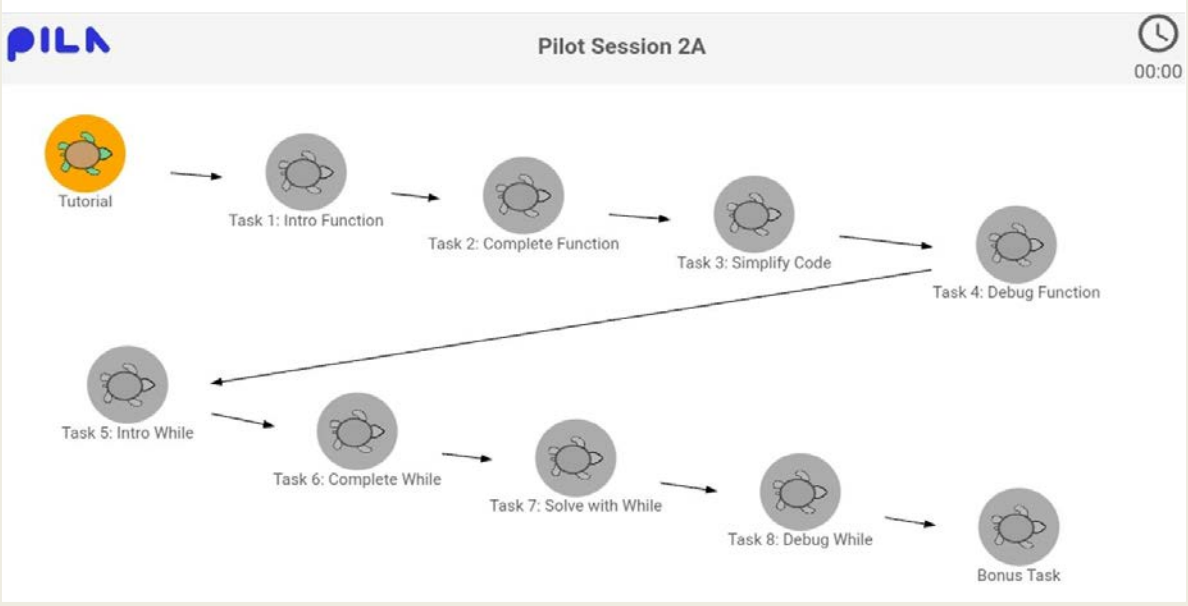
Source: OECD (n.d.[36]), OECD's Platform for Innovative Learning Assessments (PILA), https://pilaproject.org/.

Simple adaptive design for assessment experiences

Each PILA assessment experience is also structured as a progression of increasingly complex tasks (see Figure 2.3) that have a common learning target (e.g. using functions efficiently). Assessment designers and teachers have the option of locking students within a particular task until they are able to solve it (i.e. a 'level-up' mechanism) or giving students control over 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 of the assessment experience to reduce potential 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.

Figure 2.3. Assessment experience "map" (sequence of tasks) in the PILA Karel application



Source: OECD (n.d.[36]), OECD's Platform for Innovative Learning Assessments (PILA), https://pilaproject.org/.

Adaptive designs can also address the complexity of measuring learning-in-action amongst heterogenous populations of students. A relatively simple way to cater to different student ability groups 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 a "level-up" mechanism, also described in Box 2.2). More proficient students will quickly complete the initial set of simple tasks, after which they will encounter problems that challenge them. Less prepared students are also still able to engage meaningfully with the tasks and demonstrate what they can do in this task sequence design, even if they do not complete the full sequence. Both groups of students work at the cutting edge of their abilities with obvious benefits in terms of measurement quality and test engagement. With current technologies, this design could be further improved by introducing multiple adaptive paths within a scenario: on the basis of the quality of their work, students could be directed on-the-fly towards easier or more difficult sub-tasks.

Conclusion

This chapter has argued that the next generation of assessments should focus on observing and interpreting how students solve complex problems and learn to do new things. Exclusive reliance on traditional tests risks encouraging a skill-and-drill education system that does not prepare students adequately for future learning and for solving problems they have not yet seen. Students use 21st Century competencies to learn and solve problems in situations in which they interact with others, evaluate what course of action to take, consider their available resources, try out multiple strategies and adapt according to the results. If we want to assess those competencies, then we must reproduce the interactive features of these learning situations in assessments – otherwise we risk measuring something different.

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

Clearly, creating these types of assessment experiences that enable the observation of learning-in-action is not without complexity from a design perspective; making sure that they work from a measurement perspective is even more challenging. Enriching the range of assessments we use to measure students' development of 21st Century competencies will require coordinated efforts in multiple directions. First, we have to better understand how learning unfolds across different domains and types of human endeavours and how to design tasks that reproduce the key elements of the social processes behind expertise. Second, we need improved measurement models that can validly interpret complex patterns of behaviours in dynamic environments, where both a student's knowledge and the problem state change as a result of their actions. Third, we need enhanced processes to validate the inferences we make from these more complex and performance-oriented assessments, including their sensitivity to cross-cultural differences.

Finally, we need to keep in mind that the complex patterns of behaviours generated by these assessments may ultimately not be strong or separable enough to make claims about students – but it is important to find this out. We contend that we must take on this challenge because, for better or worse, assessments drive the teaching and learning that takes place within education systems. As much as teachers, business leaders and global policymakers might affirm the importance of developing competencies like persistence, critical thinking or collaboration, ultimately students and teachers will be guided by the focus of assessments.

References

Azevedo, R. and V. Aleven (eds.) (2013), *International Handbook of Metacognition and Learning Technologies*, Springer, New York, https://doi.org/10.1007/978-1-4419-5546-3. [29]

Boaler, J. and M. Staples (2008), "Creating mathematical futures through an equitable teaching approach: The case of Railside School", *Teachers College Record: The Voice of Scholarship in Education*, Vol. 110/3, pp. 608-645, https://doi.org/10.1177/016146810811000302. [9]

Chase, W. and H. Simon (1973), "Perception in chess", *Cognitive Psychology*, Vol. 4/1, pp. 55-81, https://doi.org/10.1016/0010-0285(73)90004-2. [18]

Darling-Hammond, L. et al. (2019), "Implications for educational practice of the science of learning and development", *Applied Developmental Science*, Vol. 24/2, pp. 97-140, https://doi.org/10.1080/10888691.2018.1537791. [5]

Dumont, H., D. Istance and F. Benavides (eds.) (2010), *The Nature of Learning: Using Research to Inspire Practice*, OECD Publishing, Paris, https://doi.org/10.1787/9789264086487-en. [4]

Ericsson, K. (2006), "The influence of experience and deliberate practice on the development of superior expert performance", in Ericsson, K. et al. (eds.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511816796.038. [7]

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. [31]

Hadwin, A., S. Järvelä and M. Miller (2017), "Self-regulation, co-regulation, and shared regulation in collaborative learning environments", in Schunk, D. and J. Greene (eds.), *Handbook of Self-Regulation of Learning and Performance*, Routledge, New York, https://doi.org/10.4324/9781315697048-6. [12]

Hatano, G. (1990), "The nature of everyday science: A brief introduction", *British Journal of Developmental Psychology*, Vol. 8/3, pp. 245-250, https://doi.org/10.1111/j.2044-835x.1990.tb00839.x. [11]

Hatano, G. and K. Inagaki (1986), "Two courses of expertise", in Stevenson, H., H. Azuma and K. Hakuta (eds.), *Child Development and Education in Japan*, Freeman, New York. [14]

Hatano, G. and Y. Oura (2003), "Commentary: Reconceptualizing school learning using insight from expertise research", *Educational Researcher*, Vol. 32/8, pp. 26-29, https://doi.org/10.3102/0013189x032008026. [13]

Kapur, M. and K. Bielaczyc (2012), "Designing for productive failure", *Journal of the Learning Sciences*, Vol. 21/1, pp. 45-83, https://doi.org/10.1080/10508406.2011.591717. [25]

Lei, C. and C. Chan (2012), "Scaffolding and assessing knowledge building among Chinese tertiary students using e-portfolios", in *The Future of Learning: Proceedings of the 10th International Conference of the Learning Sciences (ICLS 2012)*, International Society of the Learning Sciences, Sydney. [33]

Levy, F. and R. Murnane (2004), *The New Division of Labor: How Computers are Creating the Next Job Market*, Princeton University Press, Princeton and Oxford. [1]

Loibl, K., I. Roll and N. Rummel (2016), "Towards a theory of when and how problem solving followed by instruction supports learning", *Educational Psychology Review*, Vol. 29/4, pp. 693-715, https://doi.org/10.1007/s10648-016-9379-x. [24]

Mislevy, R. (2018), *Sociocognitive Foundations of Educational Measurement*, Routledge, New York and London. [6]

National Research Council (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, The National Academies Press, Washington, D.C., https://doi.org/10.17226/10019. [3]

Nunes, C., M. Nunes and C. Davis (2003), "Assessing the inaccessible: Metacognition and attitudes", *Assessment in Education: Principles, Policy & Practice*, Vol. 10/3, pp. 375-388, https://doi.org/10.1080/0969594032000148109. [20]

OECD (n.d.), *Platform for Innovative Learning Assessments*, https://pilaproject.org/ (accessed on 3 April 2023). [36]

Pellegrino, J. and M. Hilton (2012), *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*, The National Academies Press, Washington, D.C., https://doi.org/10.17226/13398. [2]

Piaget, J. (2001), *The Language and Thought of the Child*, Routledge, London. [8]

Roll, I. et al. (2011), "Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system", *Learning and Instruction*, Vol. 21/2, pp. 267-280, https://doi.org/10.1016/j.learninstruc.2010.07.004. [21]

Roll, I. et al. (2007), "Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking", *Metacognition and Learning*, Vol. 2/2-3, pp. 125-140, https://doi.org/10.1007/s11409-007-9010-0. [15]

Roll, I. et al. (2012), "Evaluating metacognitive scaffolding in guided invention activities", *Instructional Science*, Vol. 40/4, pp. 691-710, https://doi.org/10.1007/s11251-012-9208-7. [23]

Rosé, C. and O. Ferschke (2016), "Technology support for discussion-based learning: From computer supported collaborative learning to the future of Massive Open Online Courses", *International Journal of Artificial Intelligence in Education*, Vol. 26/2, pp. 660-678, https://doi.org/10.1007/s40593-016-0107-y. [34]

Sabatini, J., T. O'Reilly and Z. Wang (2018), "Scenario-based assessment of multiple source use", in Braasch, J., I. Braten and M. McCrudden (eds.), *Handbook of Multiple Source Use*, Routledge, New York. [19]

Schoenfeld, A. (1985), *Mathematical Problem Solving*, Academic Press, New York. [16]

Schwartz, D. and D. Arena (2013), *Measuring What Matters Most: Choice-Based Assessments for the Digital Age*, The MIT Press, https://doi.org/10.7551/mitpress/9430.001.0001. [27]

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

Shute, V., E. Hansen and R. Almond (2008), "You can't fatten a hog by weighing it - or can you? Evaluating an assessment for learning system called ACED", *International Journal of Artificial Intelligence in Education*, Vol. 18/4, pp. 289-316, https://myweb.fsu.edu/vshute/pdf/shute%202008_a.pdf (accessed on 26 March 2018). [35]

Simon, H. (1980), "Problem solving and education", in Tuma, D. and R. Reif (eds.), *Problem Solving and Education: Issues in Teaching and Research*, Erlbaum, Hillsdale, https://iiif.library.cmu.edu/file/Simon_box00013_fld00890_bdl0001_doc0001/Simon_box00013_fld00890_bdl0001_doc0001.pdf (accessed on 20 March 2023). [17]

Tan, S. et al. (2021), "Knowledge building: Aligning education with needs for knowledge creation in the digital age", *Educational Technology Research and Development*, Vol. 69/4, pp. 2243-2266, https://doi.org/10.1007/s11423-020-09914-x. [10]

Taylor, J. et al. (2010), "Using invention to change how students tackle problems", *CBE—Life Sciences Education*, Vol. 9/4, pp. 504-512, https://doi.org/10.1187/cbe.10-02-0012. [28]

van Joolingen, W. et al. (2005), "Co-Lab: Research and development of an online learning environment for collaborative scientific discovery learning", *Computers in Human Behavior*, Vol. 21/4, pp. 671-688, https://doi.org/10.1016/j.chb.2004.10.039. [30]

Vygotsky, L. (1978), *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, Massachussets. [32]

Wiske, M. (ed.) (1997), *Teaching for Understanding: Linking Research with Practice*, Jossey-Bass, San Francisco. [26]

3 Framing the focus of new assessments of 21st Century competencies

By Mario Piacentini and Natalie Foster

(OECD)

This chapter provides a simple guiding framework to identify gaps within existing assessment systems and orient decisions on which kinds of next-generation assessments to develop. The chapter does not prescribe a fixed or exhaustive list of new assessments that should be conducted nationally or internationally; rather, it provides a framing for how decision makers might map their assessment needs and define their priorities when developing new assessments. The chapter presents several examples of innovative assessments of 21st Century competencies that are consistent with this framing approach.

Introduction

As set forth in the Introduction and previous chapters of this report, mastering 21st Century competencies means being ready to successfully engage with different life situations and roles including those that are difficult to anticipate. Chapter 2 also argued that it is possible to assess these competencies by observing and interpreting how students engage with complex problems in interactive and resource-rich assessment environments that include opportunities for learning. The types of situations that students will need to successfully face throughout their lives are extremely diverse and in turn will call upon a combination of 21st Century competencies. Part of the challenge when it comes to assessing these competencies thus lies in being able to develop a sufficiently rich and diverse set of assessment experiences that adequately balances and reflects this diversity.

This chapter aims to provide some guidance on the complex challenge of determining where there are gaps in assessment systems and deciding how to address them by developing next-generation assessments (NGA). It does not prescribe a fixed or exhaustive list of assessments that should be conducted nationally and internationally; rather, it provides a simple framework to help decision makers map their assessment needs and define their priorities in the context of developing new assessments that assign value to a wider set of learning outcomes and capabilities. The chapter presents several examples of innovative assessments of 21st Century competencies that are consistent with this framing.

Towards a more comprehensive system of assessments: Framing initial design decisions

As argued in Chapter 1 of this report, different types of authentic problems or learning activities call upon a different combination of knowledge and skills. One example might be validating a claim, which primarily requires students to understand what information needs researching, to identify the available tools to conduct the research, and to engage their motivation and self-regulation skills to conduct a thorough investigation and process information critically. Another example might be debugging a dysfunctional system or inventing a more efficient method to achieve a desired outcome; these problems primarily require exploration and ideation skills, the capacity to implement and monitor strategies, and persistence.

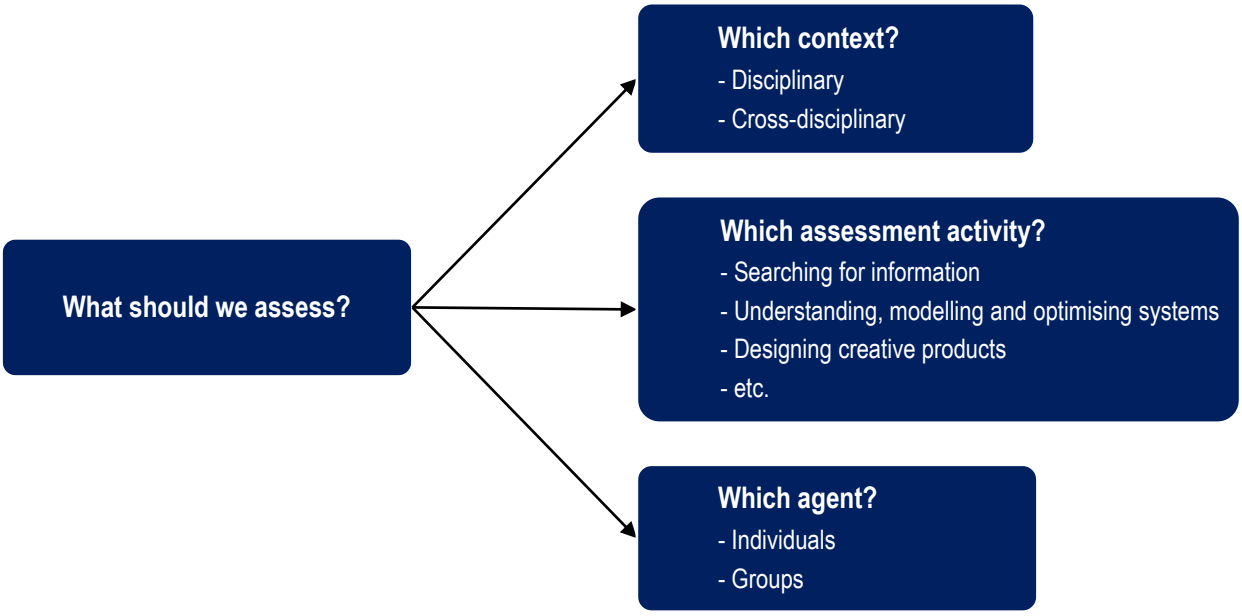
When it comes to defining the knowledge and skills required to tackle a particular problem or learning task, the context of application clearly matters too: developing an engaging movie idea and building an efficient IT network can both be considered creative activities but they each call upon a different set of domain-relevant knowledge and tools and emphasise different aspects of creativity-related skills and dispositions. Similarly, learning and problem solving in a group setting involves different processes and requires additional skills with respect to individual learning and problem solving.

The point here is to underline that decisions about what to assess and how to define assessment constructs depend on multiple contextual elements that need to be considered together and that need to be explicitly identified and addressed in the initial stages of assessment design. If the goal is to assess 21st Century competencies in a more valid and comprehensive manner, then we need a holistic system of assessments that reflect the diversity of authentic problems and learning activities that engage those competencies – in both disciplinary and cross-disciplinary contexts.

While the scope of that realisation may seem daunting, we suggest a simple framework for how we might think about determining what to assess when the intended purpose is to provide better quality information about students' development of 21st Century competencies. We argue that three interrelated questions should guide decision making on the focus of assessments (see Figure 3.1), and therefore how target constructs are defined and assessment tasks are designed:

1. *For which kinds of roles and related activities do we want to understand students' preparedness?* This question relates to explicitly defining the scenarios and assessment activity (or activities) of interest, and the relevant practices students should demonstrate while engaging in those activities.
2. *In which contexts of practice can students engage in these activities?* This question relates to acknowledging the relevant body of context-relevant knowledge, skills and attitudes that students need to engage productively with a given type of activity within a given context of practice. In other words, it relates to situating the activity within the boundaries of a discipline or making it cross-disciplinary but specifying the context(s) of application.
3. *For the purpose of assessment, are those activities organised as an individual or a group activity?* This question relates to determining whether the test taker is working independently or collaboratively during the assessment.

Figure 3.1. Framing the focus of next-generation assessments



Our perspective is that we can make robust claims on students' preparedness for their future if we collect data on how they engage with different types of relevant activities; if we contextualise those activities in a sufficiently diverse set of disciplinary and cross-disciplinary domains; and if we – for at least some of the activities – provide opportunities for collaborative work. These assessments can be designed as valuable learning activities in their own right, such that using them does not take time away from important educational targets.

Contexts of practice or domains of application

In Chapter 2 of this report, we outlined the defining features of next-generation assessments (NGA) of 21st Century competencies and argued that the types of authentic problems suited to assess them all share a set of common characteristics. These include: 1) requiring students to draw on robust knowledge schemes within a disciplinary domain or to integrate knowledge from several domains; 2) allowing students to work through them in different ways that may lead to different, yet reasonable outcomes; 3) requiring iterative processes of reasoning and doing, and therefore calling upon metacognitive reflection; and 4) enabling most students (regardless of ability level) to make some progress towards their learning and problem-

solving goal(s), while allowing proficient and motivated students to produce rich and sophisticated solutions. In sum, we suggest that NGA should be interactive and resource-rich, and based on extended, iterative and adaptive tasks that test students on their capacity to create knowledge or solve complex and authentic problems.

We also argued that a student's performance in these types of assessments depends on their knowledge or prior experience of situations that have similar patterns to the ones presented in the assessment. In other words, the domain of application clearly matters when it comes to student performance and consequently the interpretation of their performance. It is therefore critically important at the beginning of an assessment design process to make explicit and motivated choices on whether to contextualise NGA within a specific knowledge domain or to situate tasks across multiple disciplines. Cross-disciplinary here does not mean taking a domain-general approach, as the competencies that students need to engage still depend on a well-defined set of knowledge; rather, cross-disciplinary implies only that the knowledge required to engage with those tasks is not limited by the bounds of a single discipline.

Practice-oriented, disciplinary assessments

In education, the most widely used assessments of learning outcomes are typically set in one single disciplinary area (e.g. mathematics, biology, history) and primarily focus on the reproduction of acquired knowledge and procedures relevant to that discipline. We argue that, if applied to disciplinary domains along with more traditional assessment approaches, NGA could bring a better balance between the testing of disciplinary knowledge and the evaluation of students' capacity to apply this knowledge to new problems in authentic contexts.

NGAs invite students to engage in authentic practices that reflect how disciplinary knowledge is used in real life, to address both professional and everyday problems. In science, for example, an NGA 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 Chapter 4 of this report for a more detailed proposal for assessing complex problem solving in science and engineering domains). Similarly in history, students might be asked to collaboratively investigate and find biases in an historical account of an event.

New cross-disciplinary assessments

While we argue that the NGA principles outlined in Chapter 2 of this report should be adopted in subject-specific assessments to deepen evaluations of disciplinary expertise, we also acknowledge that there is value in situating NGA in contexts that do not reflect established school subjects. This is simply because life outside of school is not neatly organised by subject area. Moreover, assessment tasks that span across different disciplinary areas may better reflect the nature of classroom-based, knowledge-building activities that have proven effective for developing 21st Century competencies.

Research shows that engaging students in "epistemic games", where they simulate the work of professionals, is a productive pedagogical choice for developing 21st Century competencies (Shaffer et al., 2009[1]). Epistemic games are immersive, technology-enhanced role-playing games where players learn to think like doctors, lawyers, engineers, architects, etc. Evidence shows that what students learn during epistemic games transfers beyond the game scenario and helps them to become fluent in valuable social practices that cross disciplinary boundaries. For example, when learning to investigate a news report as journalists do, learners acquire critical thinking skills that will be valuable in their lives no matter their future profession. Many professional roles also require individuals to connect different areas of disciplinary knowledge and use their intrapersonal and interpersonal skills to solve unfamiliar problems. Some existing innovative assessments adopt this approach and simulate or immerse students in a professional task where they have to draw upon and creatively apply their knowledge from multiple disciplines (see Box 3.1).

Box 3.1. Simulating the role of a city planner in SimCityEdu

In SimCityEdu, students assume the role of a city planner (Mislevy et al., 2014[2]). The 3-D simulation environment includes many of the features of a real city: students can build or demolish buildings, transport systems, power plants and schools, as well as interrogate virtual agents on the state of the city. The assessment mission requires students to replace large polluters in the city with green technologies to improve the air quality while at the same time supporting the city's employment. The assessment within the game was designed to measure students' capacity to understand, intervene and optimise systems that are comprised of multiple variables – a 21st Century skill that is generally referred to as systems thinking or complex problem solving (Arndt, 2006[3]).

Success in the assessment demands strong metacognitive and reflective capabilities: in the fast-moving simulation, students have to analyse what happened, why it happened, what the consequences are, and what to do next. Students must also understand a city planner's role, be able to read data on employment and pollution as well as understand written language well enough to make sense of the games' goals, the citizens' complaints and other live feedback. The example nicely illustrates how simulated professional contexts provide fertile grounds for assessing 21st Century competencies that cross disciplinary domains. It also demonstrates how any assessment of a particular 21st Century skill – in this case, systems thinking – cannot be decoupled from students' prior knowledge and other skills

Figure 3.2. Screenshot from the SimCityEdu interface

Source: Mislevy et al. (2014[2]).

Simulating professional work thus represents a fertile ground for cross-disciplinary assessments. However, readiness for life – the core aspiration behind efforts to teach and measure students' development of 21st Century competencies – implies more than just being prepared for undertaking a particular job or shifting to a new job when circumstances change. Preparing students for life also involves giving them the means to make decisions in the social and civic sphere. Making responsible decisions in these contexts requires students to develop an understanding of how various social and civic problems arise as well as how they can be solved. Arguably, multiple disciplines contribute to such a complex body of knowledge.

Research shows that the capacity to act as members of a social group and in the public sphere can be acquired through collaborative, project-based learning experiences where students participate in the life of a real or simulated community, expressing opinions on issues, practicing decision making and taking on different roles (Feldman et al., 2007[4]). It follows that some NGA should be contextualised in situations where students 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 techniques, affording opportunities to make social choices and develop empathetic understanding by projecting oneself through an avatar (Raphael et al., 2009[5]). These types of simulations are particularly suited to assess socio-emotional skills like communication, cooperation and empathy, and an increasing number of role-play games have been designed to assess them in a stealth way. For example in the game Hall of Heroes, students enrol in a superhero middle school where they must develop their powers and skills to make friends, resist peer pressure and save the school from a supervillain (Irava et al., 2019[6]).

One significant challenge concerning cross-disciplinary assessments of 21st Century competencies is the lack of solid theories about how knowledge and skills develop in these contexts. Precisely defining which factors should be considered construct-relevant or -irrelevant, and what exactly constitutes "good performance", in cross-culturally valid ways, are related challenges (see Chapter 6 of this report for an in-depth example of domain analysis and modelling for complex constructs).

Relevant activities for assessing 21st Century competencies

Not all assessment problems can give us a rich body of evidence on learners' cognitive, metacognitive, attitudinal and socio-emotional skills. Traditional models of problem solving, known as phase models (Bransford and Stein, 1984[7]) suggest that all problems can be solved if individuals: 1) identify the problem; 2) generate alternative solutions; 3) evaluate those solutions; 4) choose and implement a solution; and 5) evaluate the effectiveness of the chosen solution. While these descriptions of general processes are useful, they might wrongly imply that problem solving is a uniform activity (Jonassen and Hung, 2008[8]). In reality, problems significantly vary in several important ways including the context in which they occur, their structure or openness, and the combination of knowledge and skills that the problem solver needs in order to reach a successful outcome.

We illustrate here three clusters of assessment activities that can likely provide valid evidence about how well learning experiences have prepared students for their future. We name these three clusters as follows: 1) searching for, evaluating and sharing information; 2) understanding, modelling and optimising systems; and 3) designing creative products. This is not by any means an exhaustive typology of NGA activities; a definitive typology does not exist and might even have the perverse effect of constraining the ideation of new assessments, not to mention the fact that the types of problems and activities for which students will need to be prepared for their future will continue to evolve over time. Moreover, we acknowledge that these three clusters of activities are not mutually exclusive and that they can overlap to some extent. This is particularly so in the context of classroom assessment where it would be possible to combine them into rich, extended experiences of knowledge creation and problem solving. However, we find these three clusters illustrative of different problem types that draw upon distinctly different sets of knowledge and skills but that each provide opportunities to observe higher-order thinking and learning processes.

Activity cluster 1: Searching for, evaluating and sharing information

In this cluster of activities, the main problem solving or learning goal for test takers consists of searching for and using information to reason about a problem and communicate a supported conclusion. These activities focus on how students interact with various types of media and information resources, and represent authentic and relevant activities in virtually any disciplinary or cross-disciplinary knowledge area (i.e. context of practice).

Focus competencies

Engaging in these type of activities is frequently defined in the literature as information problem solving (Brand-Gruwel, Wopereis and Vermetten, 2005[9]; Wolf, Brush and Saye, 2003[10]). Information problem solving emphasises 21st Century competencies such as critical thinking, synthesis and argumentation, communication and self-regulated learning, which are all core competencies that students need for their future – and increasingly so as the rise of the Internet and social media has brought about changes to the way people consume news and information (Brashier and Marsh, 2020[11]; Flanagin, 2017[12]). However, research shows that many students are not able to solve information problems successfully (Bilal, 2000[13]; Large and Beheshti, 2000[14]).

Selected assessment examples

In an NGA context, the sequencing of tasks in this activity cluster should require students to identify their information needs, locate information sources in online or offline environments, extract, organise and compare information from various sources, reconcile conflicts in the information, and make decisions about what information to share with others and how.

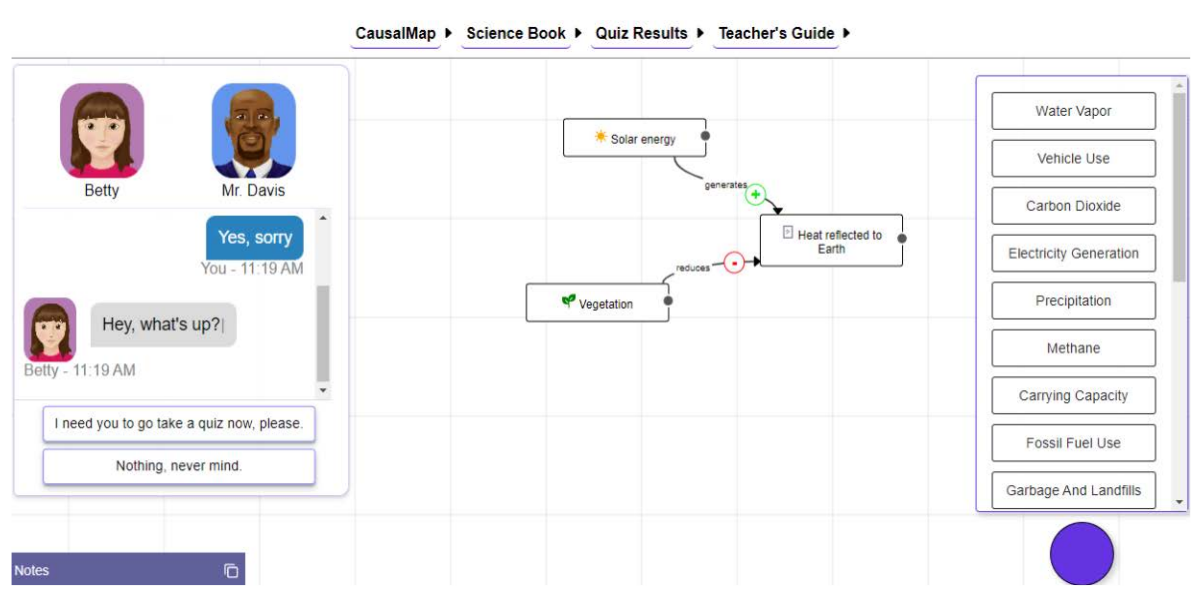
Several existing assessment examples focus on these type of information problems and incorporate the NGA design principles described in Chapter 2 of this report. In the United States, the National Assessment of Educational Progress (NAEP) Survey Assessments Innovation Lab (SAIL) "Virtual World for Online Inquiry" project (Coiro et al., 2019[15]) developed a virtual platform simulating a micro-city world, where students are presented with an open inquiry challenge (e.g. to find out whether an historical artefact should be displayed in the local museum). Students must 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 several adaptive design features (e.g. hints, prompts and levelling) to help students regulate their inquiry processes and to encourage efficient and effective information gathering.

Similar assessments can also be fully integrated within learning experiences, with evidence about students' competencies extracted in a "stealth" way by analysing the sequences of choices students make during their inquiry processes in addition to the final outcome of their information search and synthesis (see Box 3.2 on the Betty's Brain learning environment).

Box 3.2. Betty's Brain: Searching for and representing information in an open learning environment

In the Betty's Brain environment developed by researchers at Vanderbilt University (Biswas, Segedy and Bunchongchit, 2015[16]), students engage in an extended learning-by-teaching task in which they 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. The students can also interact with another computer agent, "Mr. Davis", who helps them understand how to use the system and provides useful suggestions on effective learning strategies (e.g. how to test Betty and interpret the results of the test).

Figure 3.3. Screenshot from the Betty's Brain interface



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

The learning-by-teaching paradigm implemented through a computer-based learning environment provides a social framework that engages students and helps them learn. However, over a decade of research with Betty's Brain shows that students experience difficulties when going through the complex task of building digital representations of their understanding (Biswas, Segedy and Bunchongchit, 2015[16]). In particular, many students are unable or not motivated enough to conduct systematic tests of their concept map. In order to be successful in the environment, students must apply metacognitive strategies for setting goals, developing plans for achieving these goals, monitoring their plans as they execute them, and evaluating their progress. The data collected in this blended learning and assessment environment are thus particularly valuable for evaluating students' self-regulated learning skills.

The rapid evolution of digital and social media and new modes of interacting with these kinds of information resources also need to be reflected in NGA, meaning assessments centred on this activity cluster should also expand their focus to evaluating the transmission of (mis)information and fact-checking behaviours in open, networked information environments (Ecker et al., 2022[17]). Next-generation assessments focusing on these aspects of information problem solving might draw inspiration from existing digital games and simulations that promote the development of these skills. For example, games such as "Fake It To Make It" (Urban, Hewitt and Moore, 2018[18]), "Bad News" (Roozenbeek and van der Linden, 2018[19]) or "Go Viral!" (Basol et al., 2021[20]) teach players common techniques for promoting misinformation in the hope that this prepares them to respond to it. In "The Misinformation Game", learners can engage with posts in ecologically valid ways by choosing an engagement behaviour (with options including liking, disliking, sharing, flagging and commenting), and they are provided with dynamic feedback (i.e. changes to their own simulated follower count and credibility score) depending on how they interact with posts containing either reliable or unreliable information (van der Linden, Roozenbeek and Compton, 2020[21]).

Activity cluster 2: Understanding, modelling and optimising systems

In this cluster of activities, the main problem solving or learning goal is to model a phenomenon or engineer a desired state within a dynamic system. For example, this might involve troubleshooting a malfunctioning system, generating and testing hypotheses about faulty states, or exploring a simulated environment with the goal to produce or achieve a certain output (see the simulated city example, SimCityEdu, described in Box 3.1). In short, what unites this cluster of activities is that learners have to generate their own understanding about how a (complex) system works through their interactions and experimentation with a given set of tools and then use this understanding to achieve a particular outcome or make some kind of prediction.

Focus competencies

Engaging in these types of activities emphasises the inquiry and problem-solving practices that are the focus of modern science and technology education, such as conducting reasoned experiments, understanding systems and engineering solutions. Typically, these activities require students to plan and execute actions systematically, observe, interpret and evaluate changes resulting from their interventions, and adapt their strategies based on their observations. As such, this activity cluster works particularly well when contextualised within scientific disciplines.

However, these practices are also relevant in many real-life contexts beyond scientific disciplines and these types of problems also require significant metacognitive and self-regulated learning skills. This is in part due to the various sources of complexity that these types of problems may include related to: 1) the number of different variables; 2) the mutual dependencies between variables; 3) the role of time and developments within a system (e.g. the time between an action and observed effect); 4) transparency about the involved variables and their current values; and 5) the presence of multiple levels of analysis, with potential conflicts between levels (Dörner and Funke, 2017[22]; Wilensky and Resnick, 1999[23]). This cluster of activities can therefore also provide valuable information on how well students can address complexity and uncertainty, and how persistent and goal-oriented they are.

Selected assessment examples

In an NGA context, this cluster of activities require complex system environments with multiple variables. These environments essentially function as micro-worlds in which students can make decisions about which variables to manipulate and how, and in which the environment dynamically changes either as a function of the sequence of decisions made by students or independently of them (or both).

Since the early 1980s, researchers have developed simulations of complex problems in different contexts in order to examine learning and decision making under realistic circumstances (see work by Berry and Broadbent (1984[24]) or Fischer, Greiff and Funke (2017[25]), for example). In one micro-world developed by Omodei and Wearing (1995[26]), students played the role of a Chief Fire Officer and had to combat fires spreading in a landscape using truck and helicopter fire-fighting units. The micro-world depicted a landscape comprising forest, clearings and property, the position of initial fires, the position of fire-fighting units, and the direction and strength of the wind. The problem state of the micro-world changed both independently (e.g. as a result of changes in the wind) and as a consequence of the learners' actions. Task performance in the assessment was measured as the inverse proportion of the number of cells destroyed by fire.

More recently, many micro-worlds have been developed to assess inquiry and decision-making processes in the context of science, technology, engineering and mathematics (STEM) education. One interesting example is Inq-ITS virtual lab (Gobert et al., 2013[27]), a web-based environment in which students conduct inquiry with interactive simulations and the support of various tools. In one Inq-ITS simulation, students examine how the populations of producers, consumers and decomposers are interrelated with one another. Students are asked to stabilise the ecosystem, and in order to solve the problem they have to form a hypothesis, collect data by changing the population of a selected organism, analyse the data by examining automatically generated data tables and population graphs, and communicate the findings by completing a brief lab report. Measures of students' skills are derived from the analysis of the processes they follow while conducting their investigation.

When it comes to designing the kinds of micro-worlds described here, one challenge – particularly in the context of summative assessments – relates to the level of complexity to include within the system. Simple micro-worlds with limited affordances, such as the Micro-Dyn problems used in the 2012 PISA problem solving assessments (Fischer, Greiff and Funke, 2017[25]), do not require an extended familiarisation process for learners. They can also typically present several shorter problems to students in time-limited assessment windows as compared to more complex systems that are characterised by multiple non-linear relationships, moderating variables and rebound effects; this in turn can increase the reliability of measurement claims as well as facilitate the generalisability of those claims (as evidence is accumulated by observing how students solve different problems with different tools). However, aiming to minimise complexity is not necessarily the best approach as simple simulations might not yield sufficiently valid insights into the way students deal with complexity and uncertainty (Dörner and Funke, 2017[22]).

Activity cluster 3: Designing creative products

The third cluster of activities we discuss here are those that engage students in creative work resulting in a variety of purposeful and expressive products. These practices and resulting products can be imagined in a variety of contexts, from the engineering space (e.g. inventing a new product) to the expressive (e.g. producing a work of art or writing a poem). Creative design problems can also clearly be cross-disciplinary, and there is a growing interest in incorporating "maker settings" in education. By blurring the boundaries between disciplinary subjects, "making" activities introduce students to the expression, exploration and design processes that are central to many professional domains as well as support the development of intrapersonal skills and interpersonal skills (Martinez and Stager, 2013[28]; Blikstein, 2013[29]).

Design problems are also good examples of open and relatively unstructured problems, in the sense that they include many degrees of freedom in the problem statement (which might only consist of desired goals, rather than achieving a specific objective or outcome). This ambiguity also extends to how students' products should be evaluated because responses tend to be neither right nor wrong, only better or worse.

Focus competencies

Engaging in these types of activities requires individuals to generate, elaborate and refine their ideas, emphasising 21st Century competencies like creative and critical thinking and persistence. Because of the often ill-structured and complex nature of design problems, learners also have to engage in extensive problem structuring (Goel and Pirolli, 1992[30]). Moreover, design making is an iterative activity that is not created in a vacuum: typically, the product created by the student relates to the end goal of satisfying a "client", which in turn requires students to consider different perspectives during the design process.

Selected assessment examples

In a NGA context, these cluster of activities should monitor how students engage throughout the entire design process, from the initial phases of idea generation and formation (via prototyping) through to the completion and review of a product in response to external feedback. Ideally, these assessment activities would allow learners to move naturally between phases of active designing and more reflective review of their work, and evidence should be collected both on the final product and on the processes students engage in while developing their ideas.

To date, the majority of assessments focusing on design and creative activities have been conducted in the formative space. Some of these have developed sophisticated and multidimensional rubrics to evaluate the quality of students' final products as well as their processes of invention and self-reflection (Lindström, 2006[32]) (see Box 3.3). Performance tasks replicating authentic design processes have been much rarer in summative and large-scale assessments. Beyond the constraints of available testing time, other challenges relate to providing students with resources for engaging in creative production (e.g. physical tools) and to assigning objective scores on the quality of students' work at scale – especially if the intended use of the assessment is to compare performance across different linguistic and cultural student groups. There is, however, potential in adapting best practices that have been trialled in classroom settings to a summative context (e.g. by reducing expectations on student products or by providing them with a partially-developed product they need to finish or improve), and some performance assessments have been used successfully both at small and large scale. For example, the Assessment of Performance in Design and Technology was administered to 10 000 15-year-olds in the United Kingdom (Kimbell et al., 1991[33]).

Box 3.3. Assessing students' design work in E-Scape

Kimbell (2011[31]) validated a formative assessment activity, E-Scape, where students had to design a pill dispenser. The activity lasted about six hours and was facilitated by a teacher. In validation trials, students were organised into groups of three and each student stored their own and their groups' work in a digital portfolio. The activity proceeded as a sequence of individual and group brainstorming sessions, where students sketched prototypes on paper and modelled them using given materials.

The assessment model alternated between creation and reflection phases. Approximately one hour into the activity, learners took photos of their modelling work and reported in the portfolio what they thought was going well (or not) with their work. Group members also exchanged their work and commented on each other's portfolio. Before finishing the activity, learners recorded a 30-second videoclip explaining the features of their new product and how it met the demands of the specific user.

The final portfolio was evaluated using a comparative judgement approach. Pairs of portfolios were presented to experts and, guided by a set of criteria, they were asked to identify which of the two portfolios represented the better piece of work. This simple judgement process was then repeated many times, comparing many portfolios with many judges.

Integrating collaborative tasks in assessment

In real life, people learn and develop their skills by solving complex problems collaboratively (either face-to-face or through digital media), and collaboration itself is a key 21st Century competency. Group work is increasingly used as a pedagogical practice despite the challenges teachers face to effectively structure and moderate collaborative learning (Gillies, 2016[34]), and researchers and teachers have become increasingly aware of the positive effects that collaboration can have on students' achievement and social abilities (Baines, Blatchford and Chowne, 2007[35]; Gillies and Boyle, 2010[36]). We consider collaboration to be a third dimension in our simple framework for determining what to assess as it can be introduced for any of the cluster of activities outlined above and in any disciplinary or cross-disciplinary context. In other words, integrating affordances for collaboration can introduce additional target skills in an assessment, but any collaborative tasks must nonetheless be situated within a disciplinary or cross-disciplinary context and draw upon a set of interrelated practices depending on the cluster of activities chosen.

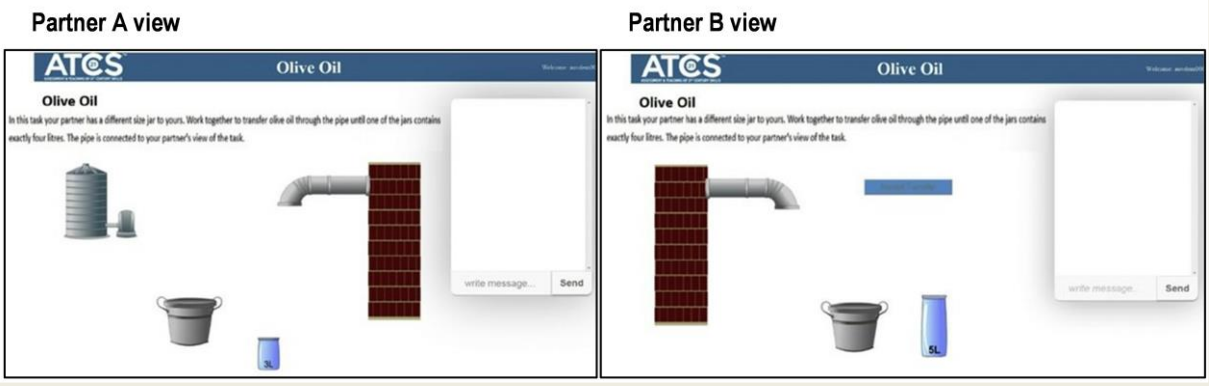
Similarly to the design activities cluster, formative assessment practices have made more progress with respect to assessing student collaboration (e.g. with teachers using rubrics to make judgements). However, two notable large-scale exceptions are the PISA 2015 collaborative problem-solving assessment and the Assessment and Teaching of 21st Century Skills (ATC21S) project (see Box 3.4 for more).

Box 3.4. Collaborative tasks in large-scale assessments: examples from PISA and ATC21S

The Assessment and Teaching of 21st Century Skills (ATC21S) project

ATC21S used interactive logical problems with an asymmetrical distribution of tools and information between paired students to incentivise collaboration. For example, in the "Olive Oil" example task (Figure 3.4), Student A and Student B needed to work together to fill Student B's jar with 4 litres of olive oil. Each student was given different tools: Student A had a 3-litre jar, an olive oil container, an "entry" pipe and a bucket, whereas Student B had a 5-litre jar, an "exit" pipe and a bucket. Neither participant was aware of what was available to their partner before interacting with them, and both students had to explore the task space to find out they needed additional resources from their partner. Students had to identify that Student A needed to fill their jar at the container and place it under the "entry" pipe so that Student B could accept the oil from the "exit" pipe. The two students could choose when and how to communicate using the free text chatbot to establish a shared understanding and solve the problem.

Figure 3.4. Asymmetric distribution of tools and information in the "Olive Oil" task



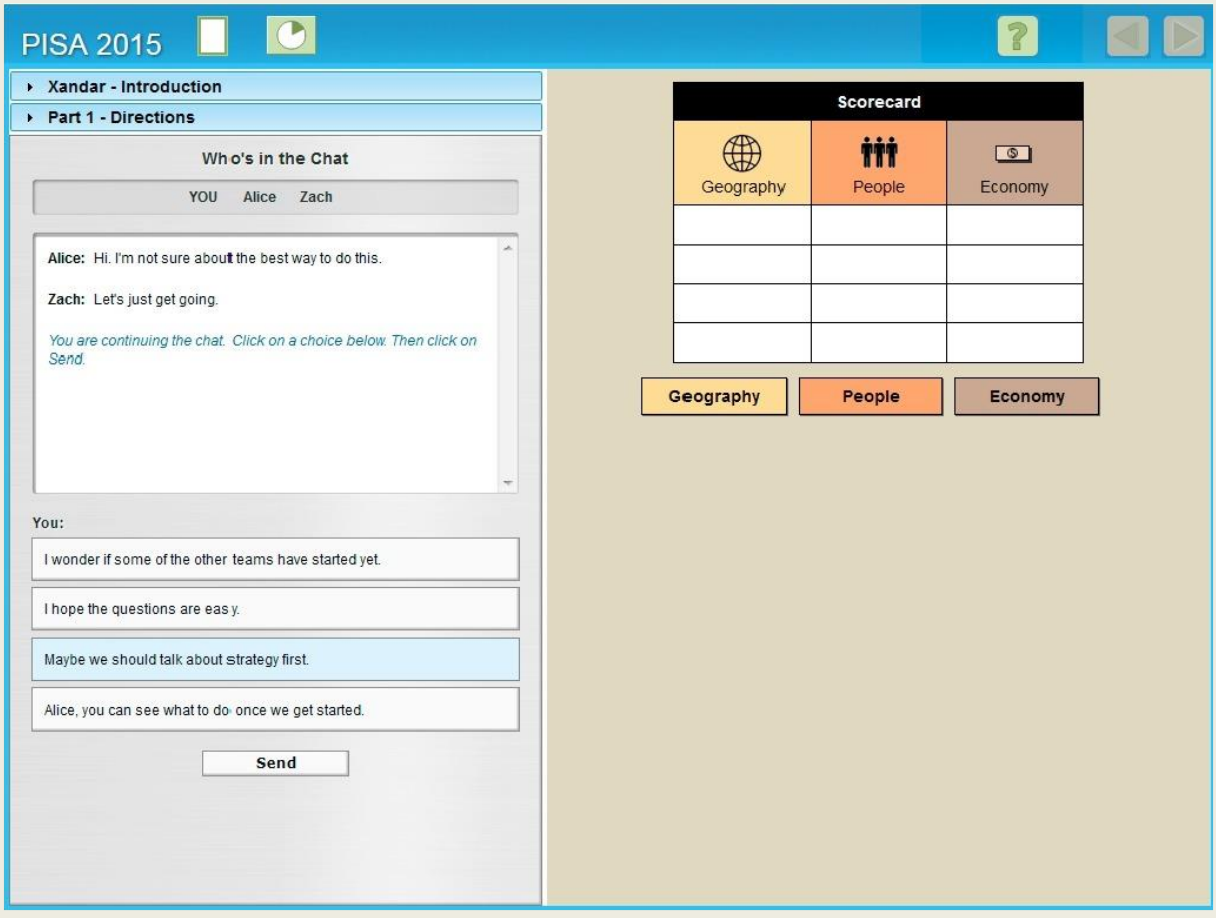
Source: Care and Griffin (2017[37]).

The PISA 2015 assessment of Collaborative Problem Solving

The PISA 2015 assessment of collaborative problem solving presented students with an interactive problem scenario and one or more "group members" (computer agents) with whom they had to interact over the course of the unit to solve the problem. Across different assessment units, the computer agent(s) were programmed to emulate different roles, attitudes and levels of competence in order to vary the type of collaborative problems students were confronted with.

The interaction of the students with the computer agent(s) were limited to pre-defined statements using a multiple-choice format and every possible intervention of the students was attached to a specific response by the computer agents or event in the problem scenario. Differing student responses could trigger different actions from the computer agents, both in terms of changes to the state of the simulation (e.g. an agent adding a piece to a puzzle) or the conversation (e.g. an agent responding to a request from the student for a piece of information).

Figure 3.5. Example task interface for the PISA 2015 Collaborative Problem-Solving Assessment



Source: OECD (2017[38]).

Although targeting similar core competencies underpinning collaboration, one major difference in approach set these two large-scale assessments apart: in PISA, students interacted with computer agents, whereas in the ATC21S assessment they engaged in real human-to-human collaboration. PISA's choice to simulate collaboration using computer agents was guided by the goals of standardising the assessment experience for students and of applying established and automated scoring methods. While the highly controlled PISA

test environment represented a trade-off in terms of the authenticity of the assessment, there are obvious challenges for scoring in tasks using human-to-human collaborative approaches. In more open assessment environments, student behaviours are difficult to anticipate and the success of any student depends on the behaviour of others in their group. This generates a measurement problem in terms of how to build separate performance scores for individual students and for student groups, and raises the issue of whether it is fair to penalise a student for the lack of ability or motivation of another student in their group.

A more recent example of a collaborative, cross-disciplinary and multi-activity cluster approach to measuring 21st Century competencies developed by the Australian Council for Educational Research (ACER) asked students to collaboratively design a plan to integrate refugees in their community (Scoular et al., 2020[39]). The first collaborative task in the assessment required students to brainstorm ideas before choosing one to implement. In the second task students were assigned group roles, with each being tied to certain responsibilities and resources, and the third task required students to bring independent research into the group discussion to improve and finalise their collaborative idea. As in the ACT21S collaborative problem-solving assessment, students exchanged ideas and information using a digital chatbot. Analysis of the chat data from multiple groups showed that it is possible to differentiate the quality of collaborative work according to different criteria, such as the level of participation of each group member or the coherence of the group's conversation. Adequate rates of agreement were reached when different raters scored the chats according to a multidimensional rubric.

Although scoring collaborative tasks at scale remains a challenge, advances in natural language processing (NLP) now make it possible to design intelligent virtual agents that "understand" what students write in open dialogues and that can respond accordingly (see Chapter 10 of this report for more on intelligent virtual agents). Providing students with choice over when and how to interact with virtual agents via open chat functions is also consistent with the vision of moving towards more choice-rich assessments and can help increase the face validity of simulated collaborative tasks by removing fixed-script constraints. Advances in NLP can also improve the quality and reduce the cost of analysing written chats among peers in genuine human-to-human collaboration, potentially enabling the automated replication of expert judgements to large sets of authentic conversational data.

One further strategy to improve inferences from conversational data is to ask students to highlight passages of their recorded conversations that they themselves consider to be evidence of good collaboration, using rubrics as a reference. These student ratings could then be used to validate the judgement of external raters or trained scoring machines and could provide in themselves additional evidence on students' understanding of what constitutes "good" communication and collaboration practices.

The pioneering assessment experiences and analytical approaches highlighted in this chapter suggest that it is possible to imagine a not-so-distant future in which collaborative tasks become an increasingly integral component of both summative and formative assessments in order to provide a more comprehensive outlook on students' development of 21st Century competencies. However, it is also clear that developing authentic collaborative tasks in next-generation assessments will require substantial parallel innovation in measurement as standard analytical models cannot yet deal with the many interdependencies across time and agents that inherently arise in collaborative settings.

Conclusion

Earlier chapters of this report established that education systems are increasingly signalling the need to develop students' 21st Century competencies. As a result, many are reforming their assessment systems with the aim of monitoring the extent to which students have developed these skills. We argue that the first step in this process should be to map current assessment gaps in order to determine which types of next-generation assessments are needed. While these decisions need to be taken by responsible actors in

each jurisdiction, according to local priorities, there is still a need for guidance from the research community on what kinds of assessment options exist, both at scale and in the classroom, and what factors are important to consider when making such decisions. This chapter offers a simple framework for guiding such decisions.

We argue that a productive way forward involves developing valid assessments of how students create knowledge and solve different types of complex problems, either on their own and collaboratively, in different contexts of application, rather than creating separate assessments for every single 21st Century skill described in the many lists that have been proposed but that abstract from authentic contexts of practice. This perspective might seem to narrow down our assessment ambitions, restricting the target to a set of skills associated with cognitive functioning – but this is not necessarily the case. Observing and interpreting how students tackle a variety of complex and contextualised problems can give us valid evidence on a wide set of cognitive, interpersonal and intrapersonal skills. Our perspective is rather to recognise one of the key conclusions from Chapters 1 and 2 of this report: that authentic problem scenarios draw upon multiple competencies simultaneously and that the context of application, the nature of the problem and the number of actors involved inherently define which combination of competencies are required for successful performance. Some clusters of activities are more similar than others in terms of the constituent elements that support performance; it follows that if we want to make claims about students' preparation for future learning, then we need to develop several different kinds of next-generation assessments that are carefully balanced in terms of their contextualisation, the purpose and organisation of the learning activity, and opportunities for collaboration.

The examples discussed in this chapter illustrate three clusters of learning and problem-solving activities that, between them, invite students to think critically and creatively, monitor their emerging understanding, preserve their motivation and goal-orientation, and regulate their learning processes and emotions. All three clusters can be situated in disciplinary or cross-disciplinary contexts. If designed to allow for multiple students working on the same problem, it would also be possible to observe how students engage in important communication and collaboration skills. Despite their promise, however, we acknowledge that many of the examples presented in this chapter have not left the lab where they were invented; and even in cases where they have gained high research visibility, they have not (yet) changed the way that assessment is done at scale. One clear implication of that is a need for more substantial investment in assessment design and validation to bring these innovation efforts to full maturity and to scale them up when they prove they can effectively measure what is hard but nonetheless important to measure.

References

Arndt, H. (2006), "Enhancing system thinking in education using system dynamics", *SIMULATION*, Vol. 82/11, pp. 795-806, https://doi.org/10.1177/0037549706075250. [3]

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. [35]

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, https://doi.org/10.1177/20539517211013868. [20]

Berry, D. and D. Broadbent (1984), "On the relationship between task performance and associated verbalizable knowledge", *The Quarterly Journal of Experimental Psychology Section A*, Vol. 36/2, pp. 209-231, https://doi.org/10.1080/14640748408402156. [24]

Bilal, D. (2000), "Children's use of the Yahooligans! 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. [13]

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. [16]

Blikstein, P. (2013), "Digital fabrication and "making" in education: The democratization of invention", in Büching, C. and J. Walter-Herrmann (eds.), *FabLab: Of Machines, Makers and Inventors*, Transcript Publishers, Bielefeld. [29]

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. [9]

Bransford, J. and B. Stein (1984), *The Ideal Problem Solver: A Guide for Improving Thinking, Learning, and Creativity*, Freeman, New York. [7]

Brashier, N. and E. Marsh (2020), "Judging truth", *Annual Review of Psychology*, Vol. 71/1, pp. 499-515, https://doi.org/10.1146/annurev-psych-010419-050807. [11]

Care, E. and P. Griffin (2017), "Assessment of collaborative problem-solving processes", in Csapó, B. and J. Funke (eds.), *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264273955-16-en. [37]

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. [15]

Dörner, D. and J. Funke (2017), "Complex problem solving: What it is and what it is not", *Frontiers in Psychology*, Vol. 8/1153, pp. 1-11, https://doi.org/10.3389/fpsyg.2017.01153. [22]

Ecker, U. et al. (2022), "The psychological drivers of misinformation belief and its resistance to correction", *Nature Reviews Psychology*, Vol. 1/1, pp. 13-29, https://doi.org/10.1038/s44159-021-00006-y. [17]

Feldman, L. et al. (2007), "Identifying best practices in civic education: Lessons from the Student Voices Program", *American Journal of Education*, Vol. 114/1, pp. 75-100, https://doi.org/10.1086/520692. [4]

Fischer, A., S. Greiff and J. Funke (2017), "The history of complex problem solving", in Csapó, B. and J. Funke (eds.), *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264273955-9-en. [25]

Flanagin, A. (2017), "Online social influence and the convergence of mass and interpersonal communication", *Human Communication Research*, Vol. 43/4, pp. 450-463, https://doi.org/10.1111/hcre.12116. [12]

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. [34]

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. [36]

Gobert, J. et al. (2013), "From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining", *Journal of the Learning Sciences*, Vol. 22/4, pp. 521-563, https://doi.org/10.1080/10508406.2013.837391. [27]

Goel, V. and P. Pirolli (1992), "The structure of Design Problem Spaces", *Cognitive Science*, Vol. 16/3, pp. 395-429, https://doi.org/10.1207/s15516709cog1603_3. [30]

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. [6]

Jonassen, D. and W. Hung (2008), "All problems are not equal: Implications for problem-based learning", *Interdisciplinary Journal of Problem-Based Learning*, Vol. 2/2, pp. 6-28, https://doi.org/10.7771/1541-5015.1080. [8]

Kimbell, R. (2011), "Evolving project e-scape for national assessment", *International Journal of Technology and Design Education*, Vol. 22/2, pp. 135-155, https://doi.org/10.1007/s10798-011-9190-4. [31]

Kimbell, R. et al. (1991), *The Assessment of Performance in Design and Technology*, School Examination and Assessment Council, London. [33]

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. [14]

Lindström, L. (2006), "Creativity: What is it? Can you assess it? Can it be taught?", *International Journal of Art & Design Education*, Vol. 25/1, pp. 53-66, https://doi.org/10.1111/j.1476-8070.2006.00468.x. [32]

Martinez, S. and G. Stager (2013), *Invent to Learn: Making, Tinkering, and Engineering in the Classroom*, Constructing Modern Knowledge Press. [28]

Mislevy, R. et al. (2014), *Psychometric Considerations in Game-Based Assessment*, GlassLab Research, Institute of Play, http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf (accessed on 21 April 2023). [2]

OECD (2017), *PISA 2015 Results (Volume V): Collaborative Problem Solving*, OECD Publishing, Paris, https://doi.org/10.1787/9789264285521-en. [38]

OECD (n.d.), *Platform for Innovative Learning Assessments*, https://pilaproject.org/ (accessed on 3 April 2023). [40]

Omodei, M. and A. Wearing (1995), "The Fire Chief microworld generating program: An illustration of computer-simulated microworlds as an experimental paradigm for studying complex decision-making behavior", *Behavior Research Methods, Instruments, & Computers*, Vol. 27/3, pp. 303-316, https://doi.org/10.3758/bf03200423. [26]

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. [5]

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. [19]

Scoular, C. et al. (2020), *Collaboration: Skill Development Framework*, Australian Council for Educational Research, Camberwell, https://research.acer.edu.au/ar_misc/42. [39]

Shaffer, D. et al. (2009), "Epistemic Network Analysis: A prototype for 21st century assessment of learning", *International Journal of Learning and Media*, Vol. 1/2, pp. 1-22, https://doi.org/10.1162/ijlm.2009.0013. [1]

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. [18]

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. [21]

Wilensky, U. and M. Resnick (1999), "Thinking in levels: A dynamic systems approach to making sense of the world", *Journal of Science Education and Technology*, Vol. 8/1, pp. 3-19, https://doi.org/10.1023/a:1009421303064. [23]

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. [10]

4 Assessing complex problem-solving skills through the lens of decision making

By Carl Wieman and Argenta Price

(Stanford University)

Solving novel complex problems is a core competency in the modern world but progress on how to effectively teach and assess such skills has lagged. This chapter argues that better characterisation of problem solving will allow for more successful assessment and teaching of this competency. The chapter presents a detailed characterisation of the problem-solving process in science and engineering domains, in which a set of core decisions that frame the problem-solving process have been empirically identified. It also presents a template for assessment designs that allow for the observation of specific decision-making processes and that address assessment design issues including the criteria for selecting good problems, integrating scaffolding for decision-making opportunities and the use of technology to collect evidence on students' processes.

Introduction: Scientific problem solving as a core competency

This chapter elaborates on one of the 21st Century competencies introduced in Chapter 1 of this report. We provide an example of applying the assessment triangle (see the Introduction chapter) to define and develop assessments for the construct of complex problem solving in science and engineering (S&E). Complex problem solving, particularly in S&E fields, is a core competency of the modern world. It is the instantiation of adaptive expertise in those fields – expert scientists and engineers are not experts because they are good at following a specific procedure or technique, rather it is because they are good at applying their knowledge and technical skills to solve complex problems in their work. Thus, problem solving is what scientists and engineers do and it is the primary goal of their education and training. Many newer science standards include problem solving, or aspects of problem solving, at the core of their standards – for example, the Next Generation Science Standards (NGSS) practice of "asking questions (science) and defining problem (engineering)" (National Research Council, 2013[1]). In the past decade, the Programme for International Student Assessment (PISA), NGSS and others have worked to improve science, technology, engineering and mathematics (STEM) education by better defining the desired core competencies and measuring how well educational systems were teaching such competencies. Problem solving and its associated skills underlie most of the cognitive and process competencies that have been identified. Here we provide a more detailed characterisation of the problem-solving process in S&E and we argue why a better characterisation of problem solving (the cognition vertex of the assessment triangle) will allow more successful assessment and teaching (interpretation vertex) of these competencies. Furthermore, we introduce assessment designs that allow observation of the constituent components of problem solving (observation vertex).

A need for specificity in defining complex problem solving in S&E

A construct to be assessed needs to be carefully defined in operational terms (i.e. the cognition vertex of the assessment triangle) for assessments to be designed to collect evidence about students' performance on elements of that construct (i.e. the observation vertex). Extensive research on problem solving and expertise has been conducted over many decades (Frensch and Funke, 1995[2]; Csapó and Funke, 2017[3]; Dörner and Funke, 2017[4]; Ericsson et al., 2018[5]). Much research has focused on looking for cognitive differences between experts and novices using limited and targeted tasks (Chi, Glaser and Rees, 1981[6]; Hegarty, 1991[7]; Larkin et al., 1980[8]; McCloskey, 1983[9]; Card, Moran and Newell, 2018[10]; Kay, 1991[11]) and has revealed important novice-expert differences in ability to identify important features and use well-organised knowledge to reduce demands on working memory. Studies that challenged experts with unfamiliar problems also found that, relative to novices, they had more deliberate and reflective strategies and could better define the problem by applying their more extensive and organised knowledge base (Schoenfeld, 1985[12]; Wineburg, 1998[13]; Singh, 2002[14]). The tasks used to measure these expert-novice differences are not very authentic, so a criticism of this body of work is that it is not clear whether what has been measured is necessary or captures what makes someone an expert performer while doing their real-life jobs (Sternberg, 1995[15]).

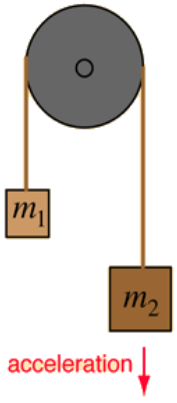
Prior assessments of problem solving have been developed both as research instruments and for standardised assessment of students. An extensive thread in this work has been to measure "domain-general" problem-solving practices (see Frensch and Funke (1995[2]) for review), but those focus on generalised practices while neglecting the necessary role that disciplinary knowledge plays in the process of solving authentic problems. Recent work on the standardised assessment of problem-solving skills has recognised the importance of the "acquisition and application of knowledge" and led to the development of more innovative assessments that do not provide all the necessary information up front to assess whether test takers recognise what information they need and how to obtain it. This work has also recognised the uncertainty involved in real problem-solving tasks (Csapó and Funke, 2017[3]). These

research and assessment efforts have contributed a better characterisation of the complexities of complex problem solving, culminating in the OECD formulation of a framework of problem solving reflecting these ideas and the development of a PISA exam to assess it (Ramalingam, Philpot and Mccrae, 2017[16]; Csapó and Funke, 2017[3]). Despite these advances, the process by which scientists and engineers specifically solve problems in their discipline has not been well studied and assessing the complex construct of problem solving in its various contexts has remained a challenge.

We have recently completed a study to identify the elements of problem solving in science and engineering, which we use to define the cognition vertex for assessment development. We developed an empirically-grounded framework that decomposes the problem-solving process into discrete cognitive components, or a set of specific decisions that need to be made during the solving process (Price et al., 2021[17]). We chose to focus on decisions because they are identifiable, measurable and important for students to practice. We identified a set of 29 decisions-to-be-made by examining the detailed problem-solving process used by experts from different areas of science and engineering. We observed that nearly all these decisions were made by every expert and that these decisions determine almost every action in the solution process. In making each of these decisions, experts invoke discipline-specific knowledge relevant to the problem's context. While the specific decisions were identified by examining the processes of experts in solving highly complex problems, most decisions apply to a large range of contexts covering nearly all educational levels. Examples of such decisions include: 1) what information is needed; 2) what concepts are relevant; 3) what is a good plan; 4) what conclusions are justified by the evidence; 5) whether the solution method works; and so on. How well these decisions are made, in a relevant context, is amenable to accurate assessment. Such assessment provides a detailed characterisation of a learner's problem-solving strengths and weaknesses including how well they can apply their relevant knowledge where needed.

Figure 4.1. *School* problems and *authentic* problems

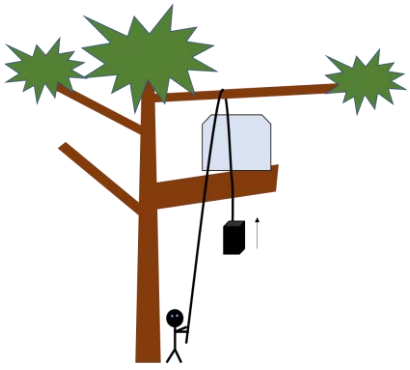
Three example physics problems of different levels of authenticity and knowledge required



Physics textbook and exam problem:
"Two masses hang from a pulley. If m_1 = 5 kg and m_2= 12 kg, calculate the acceleration of the masses. Assume the string is massless, the pulley has no mass or moment of inertia, and it is frictionless."

Photo credit: NASA

Authentic physics problem:
"Design a rocket that will launch the James Webb telescope. What will be the physical parameters of the rocket to launch the it into space and get it to its final location (size, weight, thrust, amount of fuel, etc. …)?"



Authentic but constrained physics problem:
"You are building a treehouse, using a rope hung over a branch to pull items up. How much weight can you pull up to the treehouse? How strong does the rope have to be? Is it worth the time and money to get a pulley to attach to upper branch?"

Source: Adapted from Price and Wieman (forthcoming[18]).

When thinking about the decisions involved in problem solving it is important to consider the type of problem. There is a difference between "school problems" and "authentic problems", as illustrated in

Figure 4.1. Solving the types of problems typically found on school exams and in 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 required and hence they provide very limited assessment of problem-solving skills. They do not call upon the cognitive processes required for making many of the decisions essential for solving authentic problems. Authentic problems, like the rocket example in Figure 4.1, have much greater complexity and many unknowns. Unlike the school problem, this – like all real problems – has a mixture of relevant and irrelevant information. Some of the most challenging aspects are recognising what information is needed and how to seek out that information, evaluate its reliability and apply it. In complex authentic problems, all those decisions involve incomplete information or uncertainty and so require judgement to make good decisions. These problems are sometimes referred to as "ill-structured problems" because they cannot be solved by deterministically following a set of instructions (Simon, 1973[19]). A problem can still be authentic, requiring solvers to make decisions instead of following a prescribed procedure, but be constrained to require the knowledge expected of students at a particular level (as in the tree house example in Figure 4.1). This implies that an authentic assessment of problem solving must involve problems that call upon the student to make many decisions including those that require judgement beyond just choosing to follow a memorised procedure.

Typical school problems are thus inadequate for assessing meaningful problem-solving skills and they are also poor at teaching them. In the process of solving such school problems, the student does not practice the problem-solving decisions that are necessary to thrive as a member of the STEM workforce. Research on using more complex problems in teaching, such as the Mars Mission Challenge and Jasper Woodbury problems in secondary school (Hickey et al., 1994[20]; Cognition and Technology Group at Vanderbilt, 1992[21]) and research experiences at the undergraduate level (National Academies of Sciences, Engineering and Medicine, 2017[22]), have demonstrated the benefits of having students practice complex problem solving. Assessments signal what is important for teaching, so to shift teaching toward providing practice of these more meaningful skills assessments also need to require students to make problem-solving decisions. An ideal problem could be used interchangeably as a learning activity or formative assessment in which students work through the problem in groups with timely feedback from instructors, or as summative assessment in which students solve the problem individually.

The problem-solving process of experts in science, engineering and medicine

A decision framework defined by 29 decisions

We studied the problem-solving process by systematically characterising how experts across a wide range of science (including medicine) and engineering fields solved problems in their work (Price et al., 2021[17]). In effect, we carried out a detailed domain analysis focusing on the processes involved in contextualised problem solving in S&E. From the point of view of assessment design, this domain analysis constitutes the basis of the cognition vertex of the assessment triangle. We did this through detailed interviews with over 50 experts, where they described the process of solving a specific authentic problem from their work. We then coded the interviews in terms of the decisions represented, either as explicitly stated or implied by a choice of action between alternatives. Coding these interviews required a high level of expertise in the respective discipline to recognise where choices were being made and why. We focused our analysis on what decisions needed to be made, not on the experts' rationale for making that decision: in other words, noting that a choice happened, not how they selected and chose among different alternatives for action. This is because, while the decisions-to-be-made were the same across disciplines, how the experts made those decisions varied greatly by discipline and by individual. The process of making the decisions relied on specialised disciplinary knowledge and experience. For example, planning in physics or biology may involve an extensive construction and data collection effort, while in medicine it may be running a simple blood test.

Table 4.1. Problem-solving decisions in science and engineering

Decision	Example: Ecology professor studying animal migration changes
A. Selection and goals	
1. What is important in the field?	Is/why is animal migration interesting to study?
2. Does problem fit solver's expertise?	Is my group well equipped/does my group have appropriate expertise to carry out this research?
3. What are goals, criteria, constraints?	What will be the scope of my project?
B. Frame problem	
4. Important features, concepts, info?	What environmental factors should I analyse?
5. What mental model to apply?	How are the factors that matter connected in my understanding of migration and ecological environment?
6. How to narrow down the problem? (Often involves formulating specific questions and hypotheses)	Should I look at everything that migrates? What would be a more practical subset? Are migration changes species-specific or a general trend?
7. What related problems provide useful insights?	What methods from previous meta-analyses I have conducted apply here?
8. What are potential solutions?	Is climate change affecting migration? Is migration becoming a less good strategy?
9. Is the problem solvable?	Are there enough published analyses of the type I want to consider?
C. Plan process for solving	
10. What are appropriate approximations and simplifications?	What types of publications can I not include? Can I group different species together?
11. How to decompose into sub-problems?	Should I analyse different types of animals separately? (re-address #6)
12. Most difficult or uncertain areas?	What are potential complications in my analysis?
13. What information is needed?	What publications do I need to analyse? What methods do I need to learn?
14. What are priorities in solving?	What should I start with?
15. What is the plan for getting the information needed?	What are my specific plans for analysing publications to answer my questions?
D. Interpret information and choose solutions	
16. What calculations and data analysis to perform?	What statistical methods do I use for my meta-analysis? (iterate with #15, 25)
17. How to best represent and organise info?	How should I organise results to notice patterns?
18. How believable is the info? (Valid? Reliable? What are potential biases?)	Do I take conclusions from publications at face value? How do I take data quality/quantity into account?
19. How does new data compare to predictions?	Do the factors I observe that affect migration for different species fit with my mental model? Does my mental model need updating? (iterate with #5)
20. Any significant anomalies in data?	Is there anything that does not fit?
21. What are appropriate conclusions from the data?	I'm concluding that migration is changing broadly, but for different reasons for different types of animals. How justified is that conclusion by the data? (iterate with #8, #18, #19)
22. Choose the best solution.	Does this conclusion answer my question?
E. Reflect	
23. Are the assumptions & simplifications appropriate?	As I analyse the data, does it still seem appropriate to group species and environmental factors the way I have? (re-consider #11 and #16)
24. Is additional knowledge needed?	Do I need more knowledge about statistical methods for meta-analysis? Where can I get that? Do I need more data? (iterate with #13, #15, #16)
25. How well is solving approach working?	Are my selection criteria for publications giving me the type of information I need?
26. How good is the solution?	I have limited data for some types of animals; for which do my conclusions reliably apply?
F. Implications and communication	
27. What are broader implications?	What does this mean for human impacts on animal behaviour/evolution?
28. Who is audience for communication?	Where will I publish/present this work?
29. What is the best way to present work?	How will I explain/present my work?

Notes: An illustrative example of each decision is shown, not all of the decisions made to solve the problem. Although the decisions are presented as a list, the sequence and number of iterations depends on the problem and problem solver. Decisions are not made in a sequential order and many decisions are made multiple times about different aspects of the problem or in response to or reflection about or new information.
Source: Price et al. (2021[17])

The coding identified a set of 29 decisions that experts made when solving their problems (Table 4.1). There was an unexpectedly large degree of overlap across the different fields with all experts mentioning essentially the same set of decisions. On average, each interview revealed 85% of the twenty-nine decisions and many decisions were mentioned multiple times in an interview. The set of decisions represent a remarkably consistent, yet flexible structure underlying S&E problem solving. Our interviews only show this set of decisions being made across S&E fields, but most of them are likely to apply in the social sciences and humanities as well. Research on expert thinking in history supports this view, as the thinking processes employed by expert historians align quite well with the decisions we have identified – for example, deciding what information is needed, deciding if information is believable, etc. (Wineburg, 1998[13]; Ercikan and Seixas, 2015[23]).

For the purposes of presentation, we categorised the decisions roughly based on the purposes they achieve (Table 4.1). We provide examples for each decision taken from an interview with an ecology professor studying animal migration. There are corresponding decisions from every field, but this example was relatively straightforward for a non-expert to understand. The ecologist heard a study about fish migration that piqued her interest, so she decided to conduct a meta-analysis of published work to investigate whether migration patterns are changing across animal species. In the process, she had to make all the 29 decisions identified.

The actual process is far less orderly and sequential than implied by Table 4.1 or in fact any characterisation of an orderly "scientific method." Even how problems were initiated varied widely: some experts discussed importance and goals (*decisions 1-3 – importance, fit, goals*), but others mentioned a curious observation (*decision 20 – anomalies*), important features of their system that led them to questions (*decisions 4, 6 – features, narrow*) or other starting points. We also saw that there were overlaps between decisions where two or more needed to be made together. Many decisions were also mentioned repeatedly, often about different sub-problems within the larger problem or as re-addressing the same decision in response to reflection as new information and insights were developed. The sequence and number of iterations described varied dramatically by interview, making it clear that there was no coherent "scientific process" beyond that nearly all these decisions are made at some point.

A particular feature of all the problem-solving decisions is that there is not sufficient information to know what to do with certainty. If there were, then it would be just a procedure with a clearly defined set of steps to follow and thus could be carried out by a computer. For the problem-solving decisions we have identified there is not complete information, but there is sufficient information to allow a "well-educated guess" as to what would be the best choice between options. Having expertise in the discipline then means being able to use past research and experience in the discipline to make choices that are most likely to provide a desired outcome. This definition of expertise provides a standard by which to measure the quality of an assessment question that asks the student to "make and justify" any one of these decisions.

The role, use, organisation and application of knowledge in decision making

The decisions-to-be made were consistent across all S&E disciplines we studied and most likely apply in other fields as well. However, we do not believe these decisions can be measured in a domain-general context, because how the decisions were made (and the decision outcome) was completely intertwined with the discipline and the problem. Making any of the decisions requires the application of relevant disciplinary knowledge including recognising when one does not have sufficient knowledge and/or information and so needs to seek it out.

One aspect of knowledge was common across all interviews: experts had their disciplinary knowledge organised in a manner optimised for making problem-solving decisions. Studies of expertise have previously observed highly interconnected knowledge structures (Egan and Greeno, 1974[24]; Klein, 2008[25]; Mosier et al., 2018[26]). We found that in this context of problem solving, these structures served as explicit tools that guided most decisions. These knowledge structures were composed of mental models

of the key features of the problem, the relationships between these features and an underlying level of mechanism that established those relationships and enabled making predictions. The models always involved some degree of simplification and approximation such that they were optimised for applying to the problem-solving decisions. The models provided a structure of knowledge and facilitated the application of that knowledge to the problem at hand, allowing experts to repeatedly run "mental simulations" to make predictions for dependencies and observables and to interpret new information. While the use of such predictive models was universal, the individual models explicitly reflected the relevant specialised knowledge, structure and standards of the discipline, which arguably largely define expertise in the discipline (Wieman, 2019[27]).

Examples of such models are: 1) in ecology, organisms with abundant food sources will continue to increase in number until limited by using up the food available or by increased predation; or 2) in physics, electric currents are electrons flowing through materials pushed along by an applied voltage, with the amount of current determined by the resistance of the material through which they are flowing and the size of the voltage.

Application of the decision framework to assessment across educational levels and contexts

We define the construct of "science problem solving" as the set of problem-solving decisions required for solving authentic problems in S&E. These decisions define much of the set of cognitive skills a student needs to practice and master "thinking scientifically" in any context. They are also relevant across a wide range of educational levels and contexts. Although these decisions were identified by studying the problem solving of high-level experts, we argue that they provide a broadly applicable framework for characterising, analysing and teaching S&E problem solving across all levels and contexts (except for decisions 1, 2 and 27 that are only relevant at high levels). The difference between educational levels is the relevant knowledge and predictive models needed to make these decisions wisely. Having insufficient knowledge does not negate the need to make the decisions (indeed, recognising when more knowledge is needed is one of the decisions), but the types of problems one could be expected to successfully solve would depend on the level of knowledge required. To assess the level of skill in a grade-level appropriate manner, the assessment problem scenario needs to be appropriate, and the knowledge required for making the decisions needs to match the educational level targeted.

The general applicability of the decisions is supported by other studies of student problem solving. In a study of first year university students solving introductory physics problems, the degree to which students followed the set of decisions in completing their solutions was well correlated with the correctness of their solutions (Burkholder et al., 2020[28]). We have also studied how secondary and post-secondary school students use a computer simulation to investigate and identify a hidden circuit (Salehi, 2018[29]; Salehi et al., 2020[30]; Wang et al., 2021[31]). We observed, through think-aloud interviews, a wide variation in students' solving abilities, matching the extent to which they correctly made decisions 3-26.

Our framework of decisions is consistent with previous work on "scientific practices" and expertise, but it is more complete, specific, empirically based and generalisable across S&E disciplines. To support this claim, in Table 4.2 we compare our decision framework with the PISA 2015 Scientific Literacy competency framework (OECD, 2015) that is aligned with our vision of contextualised complex problem solving because PISA aims to measure the real-world application of scientific knowledge rather than just knowledge recall. After each PISA competency, we gave the decision number from our decision framework (refer to Table 4.1) that captured how the competency is used in problem solving.

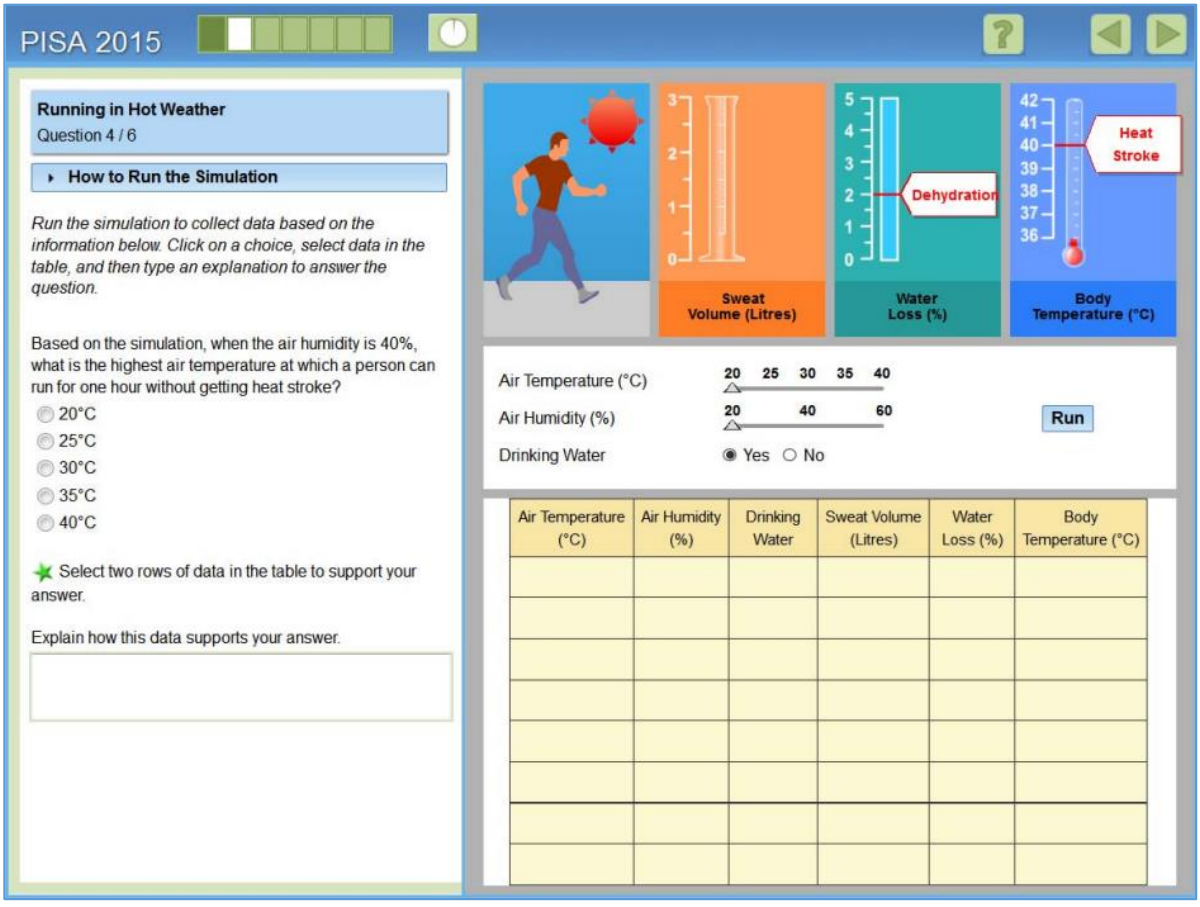
Table 4.2. Comparison of scientific literacy in PISA and problem-solving decisions in S&E

PISA 2015 scientific literacy competencies	Problem-solving decisions in S&E
1. Explain phenomena scientifically *Recognise, offer and evaluate explanations for a range of natural and technological phenomena demonstrating the ability to:*	
Recall and apply appropriate scientific knowledge	4, 5, 7
Identify, use and generate explanatory models and representations	5
Make and justify appropriate predictions	8, 16-19, 21, 22
Offer explanatory hypotheses	6
Explain the potential implications of scientific knowledge for society	27
2. Evaluate and design scientific enquiry *Describe and appraise scientific enquiries and propose ways of addressing questions scientifically demonstrating the ability to:*	
Identify the question explored in each scientific study	3, 6
Distinguish questions that are possible to investigate scientifically	9
Propose a way of exploring a given question scientifically	10-15
Evaluate ways of exploring a given question scientifically	23, 24, 25
Describe and evaluate a range of ways that scientists use to ensure the reliability of data and the objectivity and generalisability of explanations	16, 18-22
3. Interpret data and evidence scientifically *Analyse and evaluate scientific information, claims and arguments in a variety of representations and draw appropriate conclusions by demonstrating the ability to:*	
Transform data from one representation to another	17
Analyse and interpret data and draw appropriate conclusions	16, 18, 21
Identify the assumptions, evidence and reasoning in science-related texts	decisions related to this occur in next two points
Distinguish between arguments which are based on scientific evidence and theory and those based on other considerations	21, 23-26
Evaluate scientific arguments and evidence from different sources (e.g., newspaper, Internet, journals)	21, 23-26

All of the decisions in our decision framework, except decisions 1, 2, 28 and 29, occur in one or more of the PISA scientific literacy competencies and all the PISA competencies are covered by our set of decisions. However, our decision list provides greater specificity in the cognitive skills required and elements that assessment tasks should contain to collect evidence about those skills, so helps to specify what the observation vertex of the assessment triangle should include.

This can be illustrated by looking at examples of the 2015 PISA science test, such as the "Running in Hot Weather" example (Figure 4.2). This unit explores an authentic problem concerning the conditions in which it is dangerous to run and why. To solve such a problem in real life, most of the decisions on our list would be required. However, as written in the PISA example, the unit focuses primarily on two decisions: decision 15 (*plan* – in the narrow sense of whether they use a control-of-variables strategy) and decision 21 (*appropriate conclusions* – whether students correctly interpret the data they collect). The problem structure artificially narrows down and decomposes the problem and sets priorities for solving it by limiting students to consider only particular variables. The question in Figure 4.2 decides for the student that they should collect information about the effect of air temperature on body temperature specifically and even specifies how many pieces of data to collect. Essentially, the unit has made decisions 3, 4, 5, 6, 11, 13 and 14 for the students already and provides strong direction about decision 15, making it a very narrow and correspondingly incomplete measure of their scientific problem-solving ability.

Figure 4.2. "Running in Hot Weather" example problem from the PISA 2015 Scientific Literacy assessment



Source: OECD (2015[32]).

Having identified these limitations in the PISA test unit, it is straightforward to set out how to modify it to obtain a more complete assessment of problem-solving decisions. This involves a restructuring of the problem with a few additional questions to probe the student's ability to make these "missing" decisions. For example, instead of presenting a series of questions about the decomposed relationships within the simulation, students could be asked: in which conditions is it dangerous to run and why? The student could then be prompted to make specific decisions in the problem-solving process such as asking them: 1) "which variables could affect whether it is safe to run?" (*decision 4 – features*); 2) "what information do you need to collect to figure out whether it would be dangerous to go for a run on a particular day?" (*decision 13 – information needed*); and 3) "what assumptions are reasonable to make?" (*decision 10 – assumptions*). The current assessment also fixes how the data is organised – a simple alternative would be to give the student the choice of several different ways for laying out the information, thus assessing decision 17 (how to represent and organise information); then the student would be asked specific questions about interpretation of data. The simulation would have some realistic scatter in the data, as all authentic data has, and in addition to the current data interpretation questions the student would be asked about the believability of the information (decision 18). This demonstrates how the decisions framework provides a complete and more specific guide for assessing scientific competencies. This also illustrates how our set of decisions, with the three exceptions noted, applies to problem solving for almost every grade level given a suitable choice of problem context and questions.

The designers of the 2015 PISA test made choices to constrain the problem and the tested variables to make a test that was practical to score, could be completed in a few minutes and met standards of fairness in terms of specifically telling students exactly what information they needed to provide. This balance of practicality with open-endedness (to allow assessment of the full range of decisions in a problem-solving process) needs to be carefully considered in the design of any assessment. In assessments of complex thinking there is a fundamental tension between the validity of what is being measured and the ease and practicality of administering and scoring a test. The most accurate assessment of expertise would involve giving the test taker a variety of authentic problems to solve which are very broad in scope, such as "design a new cell phone" and then compare their solutions with how experts would solve the same problems – but this is obviously impractical in most circumstances. At the other extreme, an assessment question can be very easy to administer but so limited in context and scope that it involves essentially no meaningful decisions, just simple memorisation, as seen in many poor assessments. Most typical S&E assessments tend toward the "easy" end of the "validity-easy to use" spectrum. They test knowledge of information but seldom test whether the subject can correctly choose and apply knowledge in novel contexts to make good decisions to solve problems.

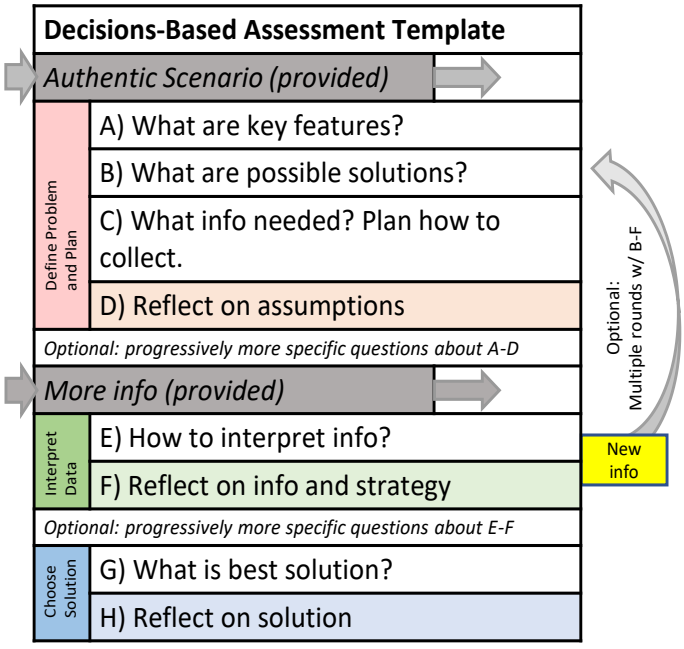
The optimum balance between authenticity and practicality is different for different level students and for different purposes of assessment (e.g. large-scale standardised tests vs. assessments within a course). Finding the appropriate balance involves choosing a problem context and questions that constrain the problem solver – but not too much – as appropriate for the assessment context. 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. Some strategies we have used for constraining the solution space, but not too much, include scaffolding the problem to probe individual decisions, including rescue points where important features are specified or information given in case students did not decide to consider those independently, and providing a flawed solution or data collection plan to be improved upon. These are discussed more in the assessment task design section below.

Designing tasks for decision-based assessments

Figure 4.3 shows a template for designing tasks (i.e. the observation vertex of the assessment triangle) to assess a large subset of problem-solving decisions described in the decision framework (i.e. the cognition vertex) (Price et al., 2022[33]). We have developed assessment tasks following this general design for undergraduate students in mechanical and chemical engineering, biochemistry, medical diagnoses and earth sciences. Problems were designed to take 30-60 minutes to complete, so are most appropriate for use in classroom or department contexts. The template can be used for assessment at lower grade levels by choosing a grade-appropriate scenario and knowledge requirements.

This assessment task starts with a realistic problem scenario that contains irrelevant and incomplete information and that to solve requires knowledge that learners are expected to know. The assessment should then call upon learners to make decisions and provide their reasoning and information used to arrive at and justify their choices. Questions on the assessment probe different decisions and the specific sequence of questions should follow the approximate order an experienced solver would use for that problem. An important part of problem solving is recognising what information is needed and correctly applying relevant information. The decisions involved in this process can be assessed by asking students to decide what information is needed, providing them with new information and then asking them to interpret and reflect upon new information.

Figure 4.3. Decisions-based assessment template



Source: Price and Wieman (forthcoming[18]), adapted from Price et al. (2022[33]).

Authentic problem scenarios are, by definition, complex, with a wide range of possible solution paths. This demands care in the design of assessment tasks to ensure the questions are sufficiently constrained so that student responses can be interpretable and readily scored, but not too constrained that the decisions are removed (as in the PISA 2015 "Running in Hot Weather" example, Figure 4.2). While each question asked to the test taker probes a different decision, certain questions will end up being interdependent by building on decisions made in previous questions. We have developed a few design strategies to build in the appropriate level of constraint and to include "rescue points" to mitigate the issue of interdependence:

- *Constrain the solution space by providing a solution* proposed by a "peer" and have test takers evaluate it and improve on its flaws. The provided "peer" work could be an answer to the problem or could be a plan for a particular sub-goal (e.g. collecting information through an experimental set-up for a biology problem). The process of troubleshooting a flawed solution or design overlaps with much of an expert problem-solving process because both involve evaluating and modifying potential solutions, including mentally testing them with new information. Thus, incorporating this kind of scenario still allows for the assessment of problem-solving decisions while significantly constraining the solution space.
- *Start with broad questions followed by questions about particularly important features.* For example, the assessment could start with questions such as "what are the important criteria for deciding when it is safe to run in hot weather?" or "what criteria will you use in evaluating the proposed design?" to avoid leading students, then follow up with rescue points in the form of questions that ask about particularly important features or that provide necessary information that students might not have collected on their own. Examples include "is sweating important for running in hot weather and if so, why?" or "how suitable was the choice of material for the design?". Learners at intermediate levels may not spontaneously recognise the significance of a factor on the initial broad question but will if prompted, allowing for a better assessment of their skills.
- *Consider the question format*. Some decisions, such as reflection decisions, need to be probed explicitly through open-ended questions. Others can be probed with complex multiple-choice

questions such as “which tests would you like to order to diagnose your patient?” followed by a list of (relevant and non-relevant) tests the student has to select or that ask for multiple-choice justifications (Walsh et al., 2019[34]). In an assessment that includes a simulation for collecting information, the assessment could also collect information about planning decisions through the actions the student takes (i.e. process data) rather than by asking the student directly. See Wang et al. (2021[31]) and Wang, Nair and Wieman (2021[35]) for work in this area.

The interpretation of data collected from an assessment (i.e. the interpretation vertex) occurs through scoring. For scoring decision-based assessments, the goal is to determine the extent to which the student: 1) decides before being explicitly prompted through built-in scaffolding; and 2) applies the correct reasoning and relevant information when making those decisions (whether prompted or not). For our assessments that probe high-level expertise, such as that desired by a student who has completed a university programme, we need to consult Ph.D. level experts to be sure of the correct information and reasoning. For problem solving assessments at lower levels, such as the PISA example, problems are less complex and require less extensive knowledge such that it is much easier to establish an “expert” decision. For most assessments, an instructor who is knowledgeable in the subject will produce an answer that is the same as someone very expert in the subject. Any question where there is not consistency among expert responses, we believe is unsuitable.

Assessment format considerations

The decision-based assessments discussed above are designed for administration in a computer-based survey format, where students fill in open-ended responses or complete complex multiple-choice questions. As test takers progress through the assessment, they are given more information and asked more questions, and are not able to go back to change earlier responses. In some cases, there is simple branching – for example, the student may only receive the additional information they request.

Simulations are an alternative format that our group has investigated. Two simulations that we have studied extensively are the PhET Interactive Simulations called “circuit construction kit black box” and the “mystery weight” (see Figure 4.4).

Figure 4.4. Examples of simulations used for assessing problem solving in S&E

Notes: On the left is the “circuit construction kit black box” problem (Salehi, 2018[29]) – students must identify the simple circuit hidden behind the black square by hooking up components and meters to the four terminals. On the right is the “mystery weight” problem (Wang, Nair and Wieman, 2021[35]) – students must determine the weight of the package by selecting weights and locating them on the beam to balance it.
Source: PhET Interactive Simulations, https://phet.colorado.edu/.

The degree of constraint and complexity is built into the simulation, thereby determining the range of decisions and problem-solving skills involved. Simulations such as these are more open-ended than the survey format, in that they allow a less constrained context in which to investigate how students decide on strategies for collecting data and how they analyse it. It is also possible to record and analyse their keystrokes ("back-end data") to get some measure of their thinking processes. The PISA 2015 simulation example is more constrained than these PhET examples, which allows for an easier interpretation of back-end data in PISA but at the expense of limiting the decisions involved.

This is work in progress, but what we have found so far is that the data collected from simulations are informative for only a limited set of decisions. While students must make many other decisions to solve the problem, it is not practical to determine these from their keystrokes (Wang et al., 2021[31]; Wang, Nair and Wieman, 2021[35]). Analysing problem solving in think-aloud interviews with these simulations, we see many of the decisions invoked in the survey format. Two unique capabilities that we have seen simulations provide is: 1) the measurement of pause time; and 2) the evolution in student strategies. Students who pause after receiving new information from the simulation, presumably to reflect on the significance of that information, perform better than students who quickly try something else with no pause (Wang, Nair and Wieman, 2021[35]; Perez et al., 2017[36]). With the simulation, we can also see that some students start with ineffective strategies but then later realise and adjust their strategies to be more effective, for example running tests that provide more useful data (Salehi, 2018[29]; Salehi et al., 2020[30]).

In conclusion, both survey and simulation format problems can sequentially provide additional information as needed. As of now, we find the survey format to be superior but further work is clearly needed in this area. With suitable affordances (e.g. what can be controlled, what can be observed, whether and how data can be collected etc.) there are likely aspects of problem-solving decisions that simulations will be better able to assess than surveys. Roll, Conati and others (Conati et al., 2015[37]; Perez et al., 2017[36]; 2018[38]) have seen how the use of different simulation designs, question prompts and analyses methods can provide other information. However, the skills and decisions that each test format is optimum for measuring and how consistent they are remain open research questions.

Conclusion

We have presented a novel framework for the assessment of complex problem solving in STEM. It is based on a set of 29 decisions that need to be made with limited information in the process of solving any authentic problem in science, whether in making choices in an individual's personal life or carrying out scientific research. The decisions were identified through a careful domain analysis by examining how scientists and engineers solved problems in their work, but we have seen how nearly all these decisions apply far more widely (across educational levels and across other disciplines). The decision framework bridges the cognition and observation vertices of the assessment triangle to provide a template for assessing the full range of knowledge and skills that it is valuable for an S&E student to learn.

An overarching implication of defining problem solving as this set of decisions is that by the time students become skilled practitioners in their fields, they will be able to make such decisions when faced with novel complex problems. This framework suggests a need for a fundamental re-evaluation of how assessments and educational experiences need to be structured to provide students with opportunities to practice making these decisions and to measure their progress toward mastery. We proposed a template for designing problems to allow practice and assessment of these decisions. A virtue of this template design is that it can be used in a very similar way for instruction as for assessment: learners would work through a problem, practicing making the various decisions (Burkholder et al., 2020[28]; Wang et al., 2022[39]). The main difference between instruction and assessment is that during instruction students would get feedback and guidance on their decisions to help them improve. This follows Ericsson's deliberate practice for the

development of expertise (Ericsson, 2006[40]). We believe there is great benefit in having good assessment and good instruction transparently connected in this fashion.

References

Burkholder, E. et al. (2020), "Template for teaching and assessment of problem solving in introductory physics", *Physical Review Physics Education Research*, Vol. 16/1, https://doi.org/10.1103/physrevphyseducres.16.010123. [28]

Card, S. (ed.) (2018), *The Psychology of Human-Computer Interaction*, CRC Press, Boca Raton, https://doi.org/10.1201/9780203736166. [10]

Chi, M., R. Glaser and E. Rees (1981), *Expertise in Problem Solving*, Learning Research and Development Center, University of Pittsburgh. [6]

Cognition and Technology Group at Vanderbilt (1992), "The Jasper Series as an example of anchored instruction: Theory, program description, and assessment data", *Educational Psychologist*, Vol. 27/3, pp. 291-315, https://doi.org/10.1207/s15326985ep2703_3. [21]

Conati, C. et al. (2015), "Comparing representations for learner models in interactive simulations", in Conati, C. et al. (eds.), *Artificial Intelligence in Education. AIED 2015. Lecture Notes in Computer Science*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-19773-9_8. [37]

Csapó, B. and J. Funke (eds.) (2017), *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, Educational Research and Innovation, OECD Publishing, Paris, https://doi.org/10.1787/9789264273955-en. [3]

Dörner, D. and J. Funke (2017), "Complex problem solving: What it is and what it is not", *Frontiers in Psychology*, Vol. 8/1153, pp. 1-11, https://doi.org/10.3389/fpsyg.2017.01153. [4]

Egan, D. and J. Greeno (1974), "Theory of rule induction: Knowledge acquired in concept learning, serial pattern learning, and problem solving", in Gregg, L. (ed.), *Knowledge and Cognition*, Erlbaum, Hillsdale. [24]

Ercikan, K. and P. Seixas (2015), "Issues in designing assessments of historical thinking", *Theory Into Practice*, Vol. 54/3, pp. 255-262, https://doi.org/10.1080/00405841.2015.1044375. [23]

Ericsson, K. (2006), "The influence of experience and deliberate practice on the development of superior expert performance", in Ericsson, K. et al. (eds.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511816796.038. [40]

Ericsson, K. et al. (eds.) (2018), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781316480748. [5]

Frensch, P. and J. Funke (1995), *Complex Problem Solving: The European Perspective*, Lawrence Erlbaum, Hillsdale. [2]

Hegarty, M. (1991), "Knowledge and processes in mechanical problem solving", in Sternberg, R. and P. Frensch (eds.), *Complex Problem Solving: Principles and Mechanisms*, Psychology Press, New York and London. [7]

Hickey, D. et al. (1994), "The Mars Mission Challenge: A generative, problem-solving School Science Environment", in Vosniadou, S., E. de Corte and H. Mandl (eds.), *Technology-Based Learning Environments*, Springer, Berlin and Heidelberg, https://doi.org/10.1007/978-3-642-79149-9_13. [20]

Kay, D. (1991), "Computer interaction: Debugging the problems", in Sternberg, R. and P. Frensch (eds.), *Complex Problem Solving: Principles and Mechanisms*, Psychology Press, New York and London. [11]

Klein, J. (2008), "Some directions for research in knowledge sharing", *Knowledge Management Research & Practice*, Vol. 6/1, pp. 41-46, https://doi.org/10.1057/palgrave.kmrp.8500159. [25]

Larkin, J. et al. (1980), "Expert and novice performance in solving physics problems", *Science*, Vol. 208/4450, pp. 1335-1342, https://doi.org/10.1126/science.208.4450.1335. [8]

McCloskey, M. (1983), "Naive theories of motion", in Gentner, D. and A. Stevens (eds.), *Mental Models*, Lawrence Erlbaum, Hillsdale. [9]

Momsen, J. (ed.) (2021), "A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment", *CBE—Life Sciences Education*, Vol. 20/3, pp. 1-15, https://doi.org/10.1187/cbe.20-12-0276. [17]

Mosier, K. et al. (2018), "Expert professional judgments and 'naturalistic decision Making'", in Ericsson, K. et al. (eds.), *The Cambridge Handbook of Expertise and Expert Performance*, Cambridge University Press, Cambridge, https://doi.org/10.1017/9781316480748.025. [26]

National Academies of Sciences, Engineering and Medicine (2017), *Undergraduate Research Experiences for STEM Students*, National Academies Press, Washington, D.C., https://doi.org/10.17226/24622. [22]

National Research Council (2013), *Next Generation Science Standards*, National Academies Press, Washington, D.C., https://doi.org/10.17226/18290. [1]

OECD (2015), *PISA 2015 Released Field Trial Cognitive Items*, https://www.oecd.org/pisa/test/PISA2015-Released-FT-Cognitive-Items.pdf (accessed on 23 March 2023). [32]

Perez, S. et al. (2017), "Identifying productive inquiry in virtual labs using sequence mining", in André, E. et al. (eds.), *Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-61425-0_24. [36]

Perez, S. et al. (2018), "Control of variables strategy across phases of inquiry in virtual labs", in Rosé, C. et al. (eds.), *Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-93846-2_50. [38]

Price, A. et al. (2022), "An accurate and practical method for assessing science and engineering problem-solving expertise", *International Journal of Science Education*, Vol. 44, pp. 2061-2084, https://doi.org/10.1080/09500693.2022.2111668. [33]

Price, A. and C. Wieman (forthcoming), "Improved teaching of science and engineering using deliberate practice of problem-solving decisions", *Innovative Teaching and Learning*. [18]

Ramalingam, D., R. Philpot and B. Mccrae (2017), "The PISA 2012 assessment of problem solving", in Csapó, B. and J. Funke (eds.), *The Nature of Problem Solving: Using Research to Inspire 21st Century Learning*, OECD Publishing, Paris, https://doi.org/10.1787/9789264273955-7-en. [16]

Salehi, S. (2018), *Improving Problem-Solving Through Reflection*, Stanford University, https://stacks.stanford.edu/file/druid:gc847wj5876/ShimaSalehi-Dissertation-augmented.pdf. [29]

Salehi, S. et al. (2020), "Can majoring in computer science improve general problem-solving skills?", *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, https://doi.org/10.1145/3328778.3366808. [30]

Schoenfeld, A. (1985), *Mathematical Problem Solving*, Academic Press, Orlando. [12]

Simon, H. (1973), "The structure of ill structured problems", *Artificial Intelligence*, Vol. 4/3-4, pp. 181-201, https://doi.org/10.1016/0004-3702(73)90011-8. [19]

Singh, C. (2002), "When physical intuition fails", *American Journal of Physics*, Vol. 70/11, pp. 1103-1109, https://doi.org/10.1119/1.1512659. [14]

Sternberg, R. (1995), "Expertise in complex problem solving: A comparison of alternative conceptions", in Frensch, P. and J. Funke (eds.), *Complex Problem Solving: The European Perspective*, Psychology Press, New York and London. [15]

Walsh, C. et al. (2019), "Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking", *Physical Review Physics Education Research*, Vol. 15/1, https://doi.org/10.1103/physrevphyseducres.15.010135. [34]

Wang, K. et al. (2022), "Measuring and teaching problem-solving practices using an interactive task in PhET simulation", in *AERA 2022 Annual Meeting*, American Educational Research Association, San Diego. [39]

Wang, K., K. Nair and C. Wieman (2021), "Examining the links between log data and reflective problem-solving practices in an interactive task", *LAK21: 11th International Learning Analytics and Knowledge Conference*, https://doi.org/10.1145/3448139.3448193. [35]

Wang, K. et al. (2021), "Automating the assessment of problem-solving practices using log data and data mining techniques", *Proceedings of the Eighth ACM Conference on Learning @ Scale*, https://doi.org/10.1145/3430895.3460127. [31]

Wieman, C. (2019), "Expertise in university teaching & the implications for teaching effectiveness, evaluation & training", *Daedalus*, Vol. 148/4, pp. 47-78, https://doi.org/10.1162/daed_a_01760. [27]

Wineburg, S. (1998), "Reading Abraham Lincoln: An expert/expert study in the interpretation of historical texts", *Cognitive Science*, Vol. 22/3, pp. 319-346, https://doi.org/10.1207/s15516709cog2203_3. [13]

Part II Innovating How We Assess

5 Exploiting technology to innovate assessment

By Natalie Foster

(OECD)

This chapter provides an overview of how digital technologies can be integrated in different assessment paradigms (traditional assessment, technology-enhanced assessment and embedded assessment). The chapter discusses the promise of technology in bridging educational rhetoric with practical assessment challenges, and in particular discusses technology-enabled innovations across the three interconnected models of an Evidence-Centred Design (ECD) assessment framework. It considers innovations at the conceptual model level (i.e. advances in the kinds of complex and dynamic performances that can be elicited), at the task model level (i.e. advances in task design) and at the measurement levels (i.e. advances in data analysis and learning analytics). The chapter underlines that technology must be used purposefully in assessment following the principles of a coherent design process.

Introduction

Digital technologies significantly expand our assessment capabilities: they offer new possibilities for designing items and test experiences, generating new potential sources of evidence, increasing efficiency and accessibility, improving test engagement and providing real-time diagnostic feedback for students and educators, among other things. These possibilities can advance assessment in several ways depending on the goals and intended purpose of assessment. Yet despite widespread recognition in the field of this transformative potential of technology for assessment, its implementation as such remains somewhat limited (Timmis et al., 2016[1]).

This chapter provides an overview of how technology can advance different assessment paradigms. It discusses innovations in three dimensions that are aligned with the three integrated models of Evidence-Centred Design (ECD). It does so to make explicit that technology does not just innovate assessment design at the task model level (i.e. how we can present assessment tasks) but also at the conceptual and measurement levels (i.e. what kinds of complex and dynamic performances we can target and elicit, and how we can generate and make sense of the data). The chapters that follow (particularly Chapters 6-10 of this report) explore some of the innovations discussed in this chapter in more detail.

The unexploited potential of technology to transform assessment

Technology offers the potential to transform what, how, why, when and where assessments happen. Several theories or models address the different uses of technology in educational contexts: some are general models of technology integration (Puentedura, 2013[2]; Hughes, Thomas and Scharber, 2006[3]), while others focus on teaching (Koehler and Mishra, 2009[4]), learning (Salomon and Perkins, 2005[5]) and assessment (Zhai et al., 2020a[6]) specifically. Despite these different perspectives, what they all share is the core idea that technology can be more or less "transformative" depending on how it is integrated into classrooms. In sum, transformative uses of technology enable new ways of learning and assessment that are otherwise not possible, providing learners with interactive tools and affordances to support their cognitive and metacognitive processes (Mhlongo, Dlamini and Khoza, 2017[7]). Less transformative uses of technology replace or functionally improve otherwise unchanged teaching, learning or assessment experiences and are often employed with the goal of making traditional activities easier, faster or more convenient.

The integration of technology in educational assessment is not new nor has its potential been completely untapped: over the past two decades, technologies have supported advances in large-scale assessment across various assessment functionalities including accessibility, test development and assembly, delivery, adaptation, adaptivity, scoring and reporting (Zenisky and Sireci, 2002[8]; Bennett, 2010[9]). Item authoring environments have expedited the production and packaging of tasks and computer-based assessment delivery has streamlined logistical processes (albeit raising other issues related to equipment, infrastructure and security). Advances in adaptive testing (i.e. presenting students with items at an appropriate level of difficulty, based on their prior responses) have also improved test efficiency and experience, and automated scoring engines have significantly reduced both the time and investment required in coder training to score assessments.

For the most part, these uses of technology have supported the computer-based delivery of otherwise unchanged tasks and item formats (i.e. selected- or constructed-response types), the automation of administrative or logistical processes, or the automation of scoring and interpreting responses (Quellmalz and Pellegrino, 2009[10]). In other words, digital technology has predominantly served to replicate or streamline existing large-scale assessment practices without significantly transforming the type of information that these assessments provide to students, teachers and other education stakeholders, nor improving the quality of test experiences for students. Such uses of technology are of course useful in

terms of efficiency gains but maintain assessment largely within the traditional assessment paradigm (i.e. drop-from-the-sky tests of accumulated knowledge). There remains considerable potential to harness digital technologies to transform what we can measure, how we make sense of test taker performance and how assessment relates to learning (Thornton, 2012[11]; Timmis et al., 2016[1]; Zhai et al., 2020a[6]; DiCerbo, Shute and Kim, 2017[12]; Shute and Kim, 2013[13])

While technology-enhanced assessment (TEA) can support more innovative assessment practices and systems, not all TEA are necessarily transformative. It is possible to identify two different TEA paradigms, in part based on the extent to which technology has been used in fundamentally transformative ways but also driven by the different goals of those assessments (Bennett, 2015[14]; Redecker and Johannessen, 2013[15]). In the first TEA paradigm, technology enables new task formats via multimedia stimuli and interactive response-types as well as improvements in measurement precision by generating new potential sources of evidence – but tests ultimately still look and feel like explicit tests for students. The other, more transformative TEA paradigm integrates assessment and learning activities through embedded assessment, where technology serves to create richer, immersive, interactive and unobtrusive learning and assessment environments that evolve as students make choices and interact with them.

Embedded assessment – also referred to as stealth assessment (Shute, 2011[16]) – engages students in carefully designed learning games and activities while unobtrusively gathering data and drawing inferences about their competencies based on their behaviours and performance. The data produced during their learning process can be used to trigger scaffolds or provide immediate feedback to learners and teachers on their progress and future learning strategies. Embedded assessments therefore minimise the gap between “teaching and learning time” and “assessment time”, translating into a fundamentally different assessment experience for learners compared to explicit testing paradigms (Redecker and Johannessen, 2013[15]).

Leveraging technology across an evidence-centred assessment design process

The remainder of this chapter discusses how technology can innovate assessment design and practice across these different assessment paradigms. Broadly speaking, the three assessment paradigms identified above can be seen as situated along a spectrum of technology integration, from least (traditional assessment) to most transformative (embedded assessment). As set forth in the Introduction chapter of this report, a principled design process like Evidence-Centred Design (ECD) (Mislevy and Haertel, 2006[17]) supports the design of valid assessments, especially when measuring complex multidimensional constructs using technology-enhanced tasks; this is because the complexity of student performances coupled with the rich observations that such tasks provide can create interpretation challenges. The following discussion is therefore structured around three key assessment design dimensions – assessment focus, task design, and evidence identification and accumulation – that closely align with the three interconnected models of an ECD framework. Table 5.1 summarises assessment capabilities across these three dimensions for each of the assessment paradigms.

Framing the discussion in this way explicitly addresses how technology enables test takers to demonstrate proficiency in new constructs, expands the ways in which tasks can be presented and behaviours observed in assessment environments, and utilises new sources of evidence and interpretative methodologies for the purposes of measurement and reporting on those constructs. While each dimension is addressed separately, there is significant interdependence among them. As the discussion advances further along the technology integration spectrum, note also that innovations enabled in less transformative uses equally apply in more transformative uses.

Table 5.1. How technology can innovate assessment

Traditional assessment (Computer-based delivery of traditional pencil-and-paper tests)	Technology-enhanced assessment (Tests with technology-enhanced stimuli and response formats)	Embedded assessment (Immersive learning and assessment experiences)
Assessment focus		
• Focus on what students already know at the time of testing (knowledge mastery and recall)	• Focus on whether students can solve well-defined tasks in interactive environments (knowledge and skills application in the context of a constrained activity)	• Focus on dynamic learning and problem-solving behaviours in open environments (knowledge and skills application in the context of an authentic performance activity)
Task design		
• Static task stimuli (text or visual) • Static problem (no learner choice or feedback) • Response types limited to multiple-choice or written constructed responses	• Interactive task stimuli (dynamic text, visual or audio) • Interactive problems and tools (some learner choice and feedback) • Wide range of response types to explicit questions	• Interactive task stimuli (dynamic text, visual or audio) or immersive virtual world • Dynamic, interactive and open-ended environments that evolve according to the students' behaviour • Wide range of possible interactions with tools and real or virtual agents
Evidence identification and accumulation		
• Limited sources of evidence • Accumulation of evidence exclusively based on number of correct and incorrect responses	• Multiple sources of evidence • Accumulation of evidence based on correct responses, solution quality and coherence of solution processes	• Multiple sources of evidence • Accumulation of evidence based on a complex, dynamic model mapping possible behaviours to measures of skill proficiency

Assessment focus

Traditional assessments typically measure what students know at a given point in time, generally by asking them to recall facts and reproduce content knowledge or to apply fixed solution procedures for static and highly structured problems (Pellegrino and Quellmalz, 2010[18]). As described in Chapter 1 of this report, these assessments cannot replicate the types of authentic problems needed to engage and capture more complex and multi-faceted aspects of performance – especially those characterised by behaviours or processes.

Technology can broaden the range of constructs we are able to measure by simulating complex problems that are open-ended and dynamic, meaning that they evolve as test takers engage in iterative processes of reasoning, problem solving and decision making. This makes it possible to measure more complex constructs or additional aspects of performance that are not possible to replicate using static tasks and problem types that focus on knowledge recall. More interactive and immersive environments allow students to engage actively in the processes of making and doing, making it possible to track the strategies that test takers employ and the decisions they make as they work through complex tasks. Taken together, these opportunities represent a powerful shift in the focus of assessment away from simple knowledge reproduction to knowledge-in-use. They also expand the range of information it is possible to capture about test takers not only in terms of what they can do (i.e. better coverage of constructs) but also in terms of how they do it (i.e. thinking and learning processes).

By simulating open-ended and dynamic environments and providing test takers with interactive tools that generate feedback, either implicitly or explicitly, test takers' knowledge and proficiency states may evolve over the course of an assessment activity – especially if it occurs over an extended period. While traditional testing models rely on assumptions about the fixed nature of knowledge and skills that constitute the target of assessment, many complex competencies explicitly include aspects of metacognition and self-regulated learning as part of their student model. Seeking and reacting to feedback is therefore an inherent part of their authentic practice (and consequently, should be part of their assessment).

Being able to track a student's knowledge and proficiency as it evolves over the course of an assessment also provides opportunities to understand how students learn and transition into higher mastery levels and how prepared they are for future learning (Xu and Davenport, 2020[19]). Technology-rich embedded assessment environments can integrate scaffolding to support learning, merging summative and formative assessment and providing measures of the capacity of test takers to learn and transfer that learning to other tasks. These types of environments are therefore particularly well-suited for both acquiring and assessing 21st Century competencies (see Chapter 2 of this report for more on 21st Century competencies and transfer of learning, and principles for designing next-generation assessments of these competencies). Research has demonstrated that these types of assessments provide evidence of latent abilities that more conventional measures fail to tap (Wolf et al., 2016[20]), more accurate measurements regarding what students know and can do (Almond et al., 2010[21]), and more useful insights and guidance to practitioners (Elliott, 2003[22]). This makes sense, given these environments are most similar to the kinds of learning contexts in which students apply and develop their skills.

While these advances enabled by technology open new possibilities for measuring complex aspects of performance, a key challenge lies in how to model these more dynamic competencies (i.e. defining how student behaviours in open environments connect to variables within the student model). Chapter 6 of this report further explores this issue.

Task design

Task design essentially refers to the tools at an assessment designer's disposal for creating valid tasks and generating relevant evidence about the target constructs. For traditional paper-and-pencil tests, this

dimension long presented a weakness to assessment design (Clarke-Midura and Dede, 2010[23]). The assessment designer's toolbox was limited to static task stimuli, multiple- or forced-choice responses among a list of pre-determined options, or simple constructed responses. As described above, these simple tasks and item-types cannot authentically replicate complex problem scenarios, nor can they generate a rich range of observations about students' mastery of 21st Century competencies – most notably about their behaviours and processes.

Digital technologies significantly expand the range of task stimuli it is possible to integrate in assessment, including text, image, video, audio, data visualisations and haptics (touch). These can enhance features of more traditional assessment stimuli (e.g. hyperlinked text or animated images) as well as create new dynamic stimuli like videos, audio and simulations. Highly interactive tools and immersive test environments can not only provide more open-ended, iterative and dynamic problems to test takers, but they can also replicate situations that would otherwise be difficult to (re)create in a standardised way for assessment (e.g. simulating collaborative encounters) or in the context of a classroom setting (e.g. visualising and modelling dynamic systems). The affordances of technology-enhanced tasks provide opportunities for test takers to make choices, iterate upon their ideas, seek feedback and create tangible representations of their knowledge and skill, and advances in technology and measurement mean that tasks can be designed to include intentional scaffolds (e.g. hints, information resources) to support learning.

New task modalities, problem types, and affordances mean there are also new possibilities for response types, and in turn, new sources of potential evidence about test takers. Digital platforms can capture, time stamp and log student interactions with the test environment and affordances. This is especially transformative in the context of measuring complex competencies as the process by which an individual engages with an activity can be just as valuable for evaluating proficiency as their final product. These process data, when coupled with appropriate analytical models, can reveal how students engage with problems, the choices they make and the strategies they do (or do not) implement – all of which may constitute potential evidence for the variables of interest in the competency model.

Chapter 7 of this report explores the expanded range of tools now at an assessment designer's disposal thanks to digital technologies, taking a deeper dive into different task formats, test affordances and response types as well as discussing how design choices must interact with validity considerations.

Evidence identification and accumulation

Test taker responses in traditional assessments are typically scored as correct or incorrect, or as polytomous items in ordered categories. In open and interactive TEA environments, test takers can provide unique, relatively complex and unstructured "responses". Potential sources of evidence may include time spent responding to a task or interacting with an affordance, sequences of actions, interactions with agents or human collaborators, multiple iterations of a product or (intermediate) solution states as well as other behaviours that can be recorded using technology (e.g. pausing, eye movements, etc.). More open-ended and interactive problem types also result in students navigating through tests in different ways. All this means that the structure and nature of the data collected can vary widely across examinees and that test items can effectively become interdependent, making it difficult or inappropriate to apply the same psychometric methods used to accumulate evidence in more traditional assessments (Quellmalz et al., 2012[24]).

Advances in measurement technology however do offer new ways to reason about and interpret evidence from TEA. Sophisticated parsing, statistical and inferential models that use Artificial Intelligence (AI), machine learning (ML) or other computational psychometric techniques can enable the automated scoring of complex constructed response data as well as accumulate evidence from different sources of data for scoring and reporting. Dynamic analytical approaches that can distinguish construct-relevant differences in student process data can integrate this information into scoring models to augment their precision. A

recent study by (Zhai et al., 2020b[25]) reviewed the technical, validity and pedagogical features of ML-involved science assessments and revealed significant advantages of these innovative assessments compared with traditional assessments. Yet exploiting these advantages requires well-designed task models that define the features of responses that matter for scoring as well as a clear conceptual understanding of how different patterns in individual student actions align with competency states (i.e. by reflecting variations in strategy or evolving psychological states). An ongoing challenge in the measurement field relates to being able to parse the massive volumes of process data generated by TEAs and to distil them into actionable pieces of information to inform claims about students' competencies (Bergner and von Davier, 2019[26]). Chapter 8 of this report focuses on this challenge and explores the potential of hybrid analytical models to reliably integrate multiple sources of evidence for scoring.

There is also potential to bridge measurement models traditionally applied to large-scale summative assessments with learning analytics methods that have been mostly applied in the context of monitoring and optimising learning and the environments in which it occurs. Learning analytics can supplement traditional performance scales by providing more descriptive and diagnostic information based on process data to explain why students might attain a given score (e.g. by identifying the strategies they successfully implemented or the types of mistakes they committed). This kind of detailed assessment reporting can help to promote a vision of assessment that is more integrated with the processes of teaching and learning. However, more work is needed to connect the disciplines of measurement science and learning analytics (see Chapter 13 of this report for directions on how to integrate advances from the two fields).

Discussion

Technology clearly offers many new possibilities for assessment design that can help to bridge the gap between educational rhetoric – the goal of assessing students' preparedness for their future – and the challenges that pose difficulties for assessing 21st Century competencies (see Chapter 1 of this report for an in-depth discussion of these challenges). Technology is also particularly suited for responding to the design innovations for next-generation assessments outlined in Chapter 2 of this report, namely providing opportunities for learning, feedback and instructional support during extended performance tasks. However, using technology to advance assessment in the ways described in this chapter warrants some further reflection.

First, technology-enabled innovations are only useful insofar as they are integrated purposefully within a principled design process. This means that choices about what aspects of performance to simulate, what tools and affordances to include, what evidence to collect and what interpretations to draw from the data are guided by an explicit chain of reasoning. Integrating more technology in assessment for technology's sake is not always better. To borrow from other frameworks on technology integration in education, in order to create meaningful learning and assessment experiences then it is necessary to bring together technology with pedagogical and content knowledge (Koehler and Mishra, 2009[4]).

Second, while the assessment paradigms described in Table 5.1 are presented along a technology integration continuum, from less to more transformative, it is worth acknowledging that they are conceptually different in terms of their intended purpose. Assessments in the first two paradigms (traditional and technology-enhanced) intend to measure a fixed state or ability, usually in the context of summative assessment; technology-enabled innovations have thus tended to focus on increasing measurement efficiency or enhancing item-types. Conversely, embedded assessments can measure dynamic abilities and processes (that may change during the assessment) and technology has been used to enable more personalised learning and assessment experiences, primarily in the context of formative assessment. Both types of assessment remain useful: if education stakeholders need a snapshot of students' knowledge across a vast content discipline (e.g. mathematics), then a more traditional assessment might serve that purpose better by more efficiently sampling activities across the domain. This

chapter does not intend to argue that one paradigm is inherently better than the other, nor that less transformative assessments will become obsolete – at least, not in the near future. However, the single-occasion, drop-from-the-sky model of assessment made of short and discrete items is clearly insufficient for measuring complex constructs that are dynamic, iterative and that require extended performance-based assessments. Measuring these constructs well necessarily requires assessments to be closer tied to the processes and contexts of learning and instruction – something that technology can facilitate through embedded assessment.

Third, the embedded assessment paradigm aligns with other visions of assessment in technology-rich environments that are characterised by performance-based formative activities (DiCerbo, Shute and Kim, 2017[12]; Shute et al., 2016[27]); see also Chapters 2 and 3 of this report. In this vision of assessment, students are immersed in different technology-rich learning environments that can capture and measure the dynamic changes in their knowledge and skills, and that information can then be used to further enhance their learning. This does not necessarily involve administering assessments more frequently, but rather unobtrusively collecting data as students learn and interact with their digital environment and systematically accumulating evidence about what students know and can do in multiple contexts – therefore merging summative and formative assessment (see (Wilson, 2018[28]), for more on facilitating coherence between classroom and large-scale assessment). However, turning this vision into reality requires the development of high-quality, ongoing, unobtrusive and technology-rich assessments whose data can be aggregated to describe a student's evolving competency levels (at various grain sizes) and also aggregated across students (e.g. from student to class, to school to district, to state to country) to inform higher-level decisions (DiCerbo, Shute and Kim, 2017[12]). This is turn implies much higher development and data analysis costs as well as addressing potential fairness issues including how such assessments are communicated to students (i.e. do they know what they are being assessed on?) and ensuring that innovative analytical and scoring models are unbiased.

Conclusion

This chapter has described how technology can innovate assessment across the three interconnected dimensions of assessment design in different assessment paradigms. Technology can introduce new forms of active, immersive and iterative performance-based tasks within interactive environments that make it possible to observe how test takers engage in complex and authentic activities. These types of tasks can provide richer observations and potential evidence about students' thinking processes and learning as well as enable the measurement of dynamic skills beyond the capability of more traditional and static items. Interactive tools and affordances can also provide dynamic and targeted feedback to test takers, supporting test takers' learning progress, motivation and engagement.

Technology can also generate new sources of complex evidence that, when combined with sophisticated analytical approaches, can identify patterns of behaviour associated with different mastery levels, increase the precision of performance scores and produce diagnostic information on what support students need in order to progress their skills. By allowing for the real-time measurement of students' capacities as they engage in meaningful learning activities, technology holds the promise of creating new systems of evaluation where evidence on students' progress is collected in a continuous way and assessment and learning are no longer explicitly separated.

While these opportunities are exciting, the development of effective innovative assessments must follow the principles of a coherent design process so that technology effectively serves the intended purposes of the assessment. The opportunities discussed in this chapter also raise a set of new challenges for assessment designers and measurement experts. These include how to design tasks that can simulate authentic contexts and elicit relevant behaviours/evidence, how to interpret and accumulate the numerous

sources of data that TEAs can create in meaningful and reliable ways, and how to compare students meaningfully in increasingly dynamic and open test environments.

References

Almond, P. et al. (2010), "Technology-enabled and universally designed assessment: Considering access in measuring the achievement of students with disabilities - A foundation for research", *The Journal of Technology, Learning and Assessment*, Vol. 10/5, http://www.jtla.org (accessed on 6 May 2022). [21]

Bennett, R. (2015), "The changing nature of educational assessment", *Review of Research in Education*, Vol. 39/1, pp. 370-407, https://doi.org/10.3102/0091732X14554179. [14]

Bennett, R. (2010), "Technology for large-scale assessment", in Peterson, P., E. Baker and B. McGaw (eds.), *International Encyclopedia of Education*, Elsevier, Oxford, https://doi.org/10.1016/B978-0-08-044894-7.00701-6. [9]

Bergner, Y. and A. von Davier (2019), "Process data in NAEP: Past, present, and future", *Journal of Educational and Behavioral Statistics*, Vol. 44/6, pp. 706-732, https://doi.org/10.3102/1076998618784700. [26]

Clarke-Midura, J. and C. Dede (2010), "Assessment, technology, and change", *Journal of Research on Technology in Education*, Vol. 42/3, pp. 309-328, https://doi.org/10.1080/15391523.2010.10782553. [23]

DiCerbo, K., V. Shute and Y. Kim (2017), "The future of assessment in technology-rich environments: Psychometric considerations", in Spector, J., B. Lockee and M. Childress (eds.), *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice, and Policy*, Springer International Publishing, New York, https://doi.org/10.1007/978-3-319-17727-4_66-1. [12]

Elliott, J. (2003), "Dynamic assessment in educational settings: Realising potential", *Educational Review*, Vol. 55, pp. 15-32, https://doi.org/10.1080/00131910303253. [22]

Hughes, J., R. Thomas and C. Scharber (2006), "Assessing technology integration: The RAT - Replacement, Amplification, and Transformation - framework", in *Proceedings of SITE 2006: Society for Information Technology & Teacher Education International Conference*, Association for the Advancement of Computing in Education, Chesapeake, http://techedges.org/wp-content/uploads/2015/11/Hughes_ScharberSITE2006.pdf. [3]

Koehler, M. and P. Mishra (2009), "What is technological pedagogical content knowledge?", *Contemporary Issues in Technology and Teacher Education*, Vol. 9/1, https://citejournal.org/volume-9/issue-1-09/general/what-is-technological-pedagogicalcontent-knowledge/. [4]

Mhlongo, S., R. Dlamini and S. Khoza (2017), "A conceptual view of ICT in a socio-constructivist classroom", in *Proceedings of the 10th Annual Pre-ICIS SIG GlobDev Workshop*, Seoul, South Korea, https://www.researchgate.net/publication/322203195. [7]

Mislevy, R. and G. Haertel (2006), "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. [17]

Pellegrino, J. and E. Quellmalz (2010), "Perspectives on the integration of technology and assessment", *Journal of Research on Technology in Education*, Vol. 43/2, pp. 119-134, https://files.eric.ed.gov/fulltext/EJ907019.pdf. [18]

Puentedura, R. (2013), *SAMR and TPCK: An Introduction*, http://www.hippasus.com/rrpweblog/archives/2013/03/28/SAMRandTPCK_AnIntroduction.pdf (accessed on 24 March 2023). [2]

Quellmalz, E. and J. Pellegrino (2009), "Technology and testing", *Science*, Vol. 323/5910, pp. 75-79, https://doi.org/10.1126/science.1168046. [10]

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.simscientists.org/downloads/Chapter_2012_Quellmalz.pdf. [24]

Redecker, C. and Ø. Johannessen (2013), "Changing assessment - towards a new assessment paradigm using ICT digital competence view project the future of learning view project", *European Journal of Education: Research, Development and Policy*, Vol. 48/1, pp. 79-96, https://doi.org/10.2307/23357047. [15]

Salomon, G. and D. Perkins (2005), "Do technologies make us smarter? Intellectual amplification with, of and through technology", in Sternberg, R. and D. Preiss (eds.), *Intelligence and Technology: The Impact of Tools on the Nature and Development of Human Abilities*, Lawrence Erlbaum, New Jersey. [5]

Shute, V. (2011), "Stealth assessment in computer-based games to support learning", in Tobias, S. and J. Fletcher (eds.), *Computer Games and Instruction*, Information Age Publishing, Charlotte, https://myweb.fsu.edu/vshute/pdf/shute%20pres_h.pdf. [16]

Shute, V. and Y. Kim (2013), "Formative and stealth assessment", in Spector, J. et al. (eds.), *Handbook of Research on Educational Communications and Technology (4th Edition)*, Lawrence Erlbaum, New York. [13]

Shute, V. et al. (2016), "Advances in the science of assessment", *Educational Assessment*, Vol. 21/1, pp. 1-27, https://doi.org/10.1080/10627197.2015.1127752. [27]

Thornton, S. (2012), "Issues and controversies associated with the use of new technologies", in *Teaching Politics and International Relations*, Palgrave Macmillan UK, London, https://doi.org/10.1057/9781137003560_8. [11]

Timmis, S. et al. (2016), "Rethinking assessment in a digital age: Opportunities, challenges and risks", *British Educational Research Journal*, Vol. 42/3, pp. 454-476, https://doi.org/10.1002/berj.3215. [1]

Wilson, M. (2018), "Making measurement important for education: The crucial role of classroom assessment", *Educational Measurement: Issues and Practice*, Vol. 37/1, pp. 5-20, https://doi.org/10.1111/emip.12188. [28]

Wolf, M. et al. (2016), "Integrating scaffolding strategies into technology-enhanced assessments of English learners: Task types and measurement models", *Educational Assessment*, Vol. 21/3, pp. 157-175, https://doi.org/10.1080/10627197.2016.1202107. [20]

Xu, L. and M. Davenport (2020), "Dynamic knowledge embedding and tracing", in Raffety, A. et al. (eds.), *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, https://files.eric.ed.gov/fulltext/ED607819.pdf. [19]

Zenisky, A. and S. Sireci (2002), “Technological innovations in large-scale assessment”, *Applied Measurement in Education*, Vol. 15/4, pp. 337-362, https://doi.org/10.1207/S15324818AME1504_02. [8]

Zhai, X. et al. (2020a), “From substitution to redefinition: A framework of machine learning-based science assessment”, *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1430-1459, https://doi.org/10.1002/tea.21658. [6]

Zhai, X. et al. (2020b), “Applying machine learning in science assessment: A systematic review”, *Studies in Science Education*, Vol. 56/1, pp. 111-151, https://doi.org/10.1080/03057267.2020.1735757. [25]

6 Defining the conceptual assessment framework for complex competencies

By Mario Piacentini

(OECD)

This chapter outlines the key elements and tools involved during the initial phases of complex assessment design, focusing on the decisions required of assessment designers to delimit the assessment domain, progressively define the assessment arguments, describe the characteristics of tasks and identify a suitable measurement model. Evidence-Centred Design (ECD) is presented as a guiding framework for making these coherent design decisions. The chapter then illustrates the process of assessment framework development for a complex construct using the example of the PISA 2025 Learning in the Digital World assessment. This example typifies the complexities of building a new assessment when the target competencies lack established definitions and theories of development and when part of the assessment validity argument relies on correctly interpreting test takers' behaviours in open and interactive simulation environments.

Introduction

Developing any new assessment is a challenging task. Assessment design essentially involves a series of decisions starting from identifying broad objectives and possibilities before arriving at specific tasks and measures, progressively adding in relevant constraints related to the context and circumstances of the intended assessment. In this process, assessment developers must answer several interconnected questions: 1) what claims do we want to make about students and how will these claims be used? 2) In which type of situations do we expect students to apply the skills that we want to make claims about? 3) What defines good performance in those situations? 4) What constitutes evidence of different skill levels? 5) Are those skills best evidenced in the processes or in the products of students' work, or do we need to consider both? 6) How can we accumulate the many observations of students' actions and behaviours into scores? 7) How do we report these scores so that assessment users can understand them and make justified conclusions? As the competencies targeted for an assessment become more complex, so does answering these questions – heightening the importance of following a principled design process that builds in validity considerations at each step along the way.

This chapter outlines the key elements and tools of the initial phases of assessment design, focusing on the decisions required of assessment designers to delimit the assessment domain and progressively define the assessment arguments, describe the characteristics of tasks and identify a suitable measurement model. The process of Evidence-Centred Design (ECD) (Mislevy and Riconscente, 2006[1]) is presented as a guiding framework for making these coherent design decisions. ECD is particularly useful for conceptualising and designing assessments of complex competencies in interactive environments, where students can acquire new knowledge and skills through automated feedback and interaction with learning resources. The chapter then illustrates the framework development process using the example of the Learning in the Digital World assessment, to be administered in the 2025 cycle of the Programme for International Student Assessment (PISA). This example was chosen as it typifies the complexities of building up a new assessment when the target competency lacks established definitions and theories of development and when part of the validity argument relies on correctly interpreting test takers' behaviours in open simulation environments.

Complexity breeds complexity: The importance of theory to orient initial design decisions

The conceptual assessment framework defines the target constructs and the key characteristics of an assessment. In general, this framework includes an analysis of the domain, describes the target constructs, lists the latent variables to be measured and describes expected progressions on these variables, defines features of families of assessment tasks, indicates how students' behaviours on these tasks can be converted into scores, and how these scores are accumulated to provide summary metrics and make claims about test takers. ECD provides a framework for connecting all these pieces of the assessment puzzle into a coherent frame, where specifications for task design, task performance and competency estimates are explicitly linked via an evidentiary chain.

In test development, arguably 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 then 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[1]). It is far more likely that an assessment achieves its intended purpose when the nature of the construct guides the construction of relevant tasks as well as the development of construct-based scoring criteria and rubrics (Messick, 1994[2]).

This critical activity becomes more challenging as the complexity of the domain (and its constituent competencies) increases. One source of difficulty stems from the lack of validated theories about the nature of such domains as well as models of how students progress in the development of relevant competencies. If we want to assess reading ability, for example, 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 of the target domain or knowledge on learning progressions is not available for more complex competencies like collaborative problem solving or communication.

Another difficulty relates to the multidimensional nature of these competencies: for example, collaborative problem solving involves both the capacity to solve problems in a given domain as well as the skills and attitudes to work effectively with others. These two sets of capabilities interact in complex ways, making it hard to disentangle them from one another when reporting on students' performance in collaborative tasks.

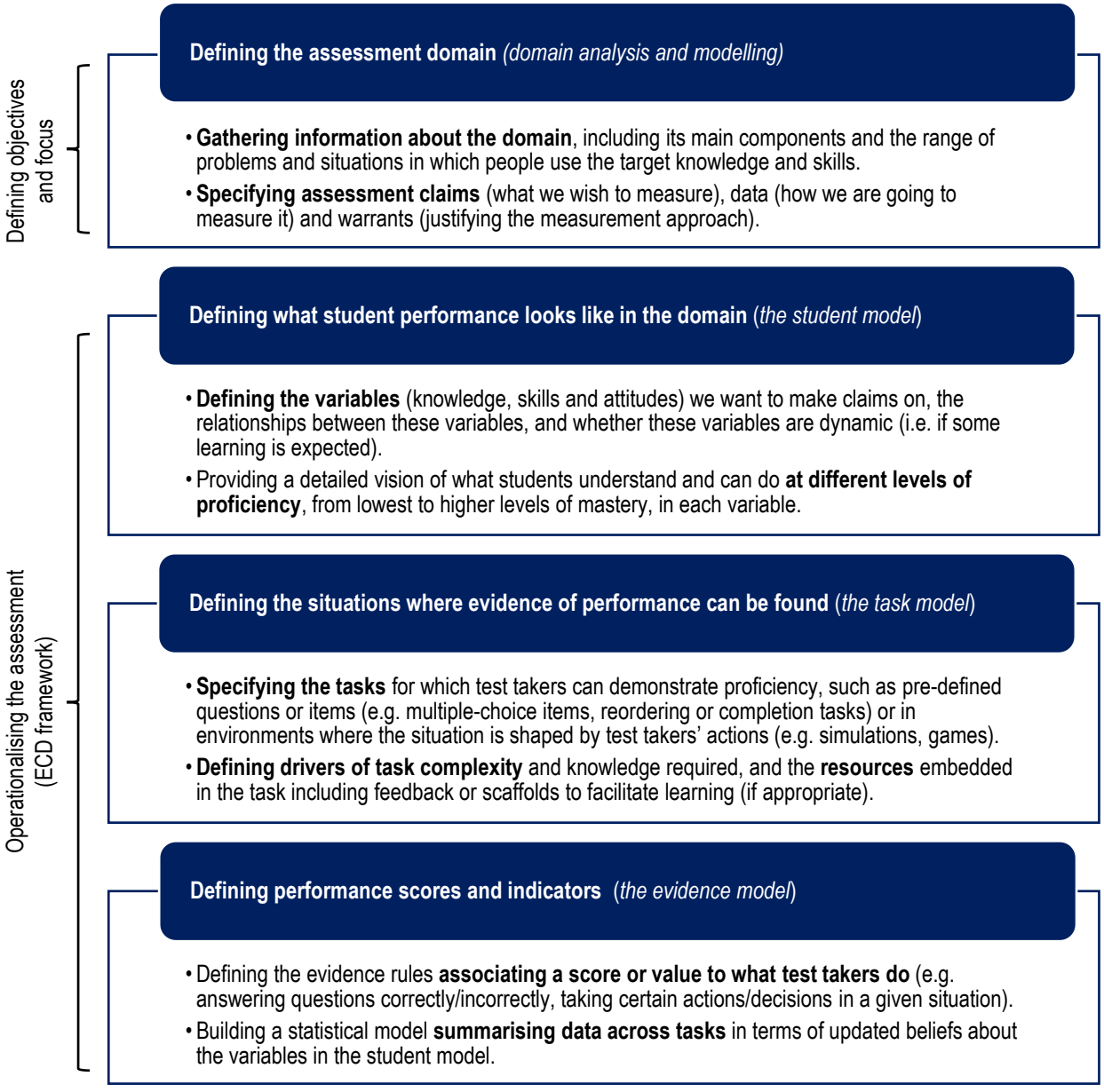
A third difficulty arises from the way that complex competencies manifest themselves in the real world. As discussed in previous chapters of this report, validly assessing these competencies requires being able to replicate complex performance tasks from which we can observe how people behave and adapt to problem situations. Identifying those situations, reproducing their core features in authentic digital simulations and translating traces of actions within the digital test environment into evidence for the claims represents a sequence of hurdles that assessment designers have to overcome.

In short, every new assessment needs to be guided by a theory of learning in the domain that identifies what is important to measure. Anchoring the subsequent design of tasks and the evidence model into a well-defined theoretical framework is therefore essential for generating valid inferences about test takers' performance. However, while the theoretical knowledge accumulated during the domain analysis phase provides critical direction, especially in the initial stages of an assessment design process, it does not closely prescribe what the final assessment will look like.

Each assessment is ultimately the result of a creative design process, where a design team strives to achieve near-optimal solutions under multiple (sometimes conflicting) constraints and a certain degree of uncertainty. This design process is also iterative: several rounds of revisions are typical and, more importantly, necessary when developing assessments of complex competencies that make use of extended performance tasks. Test takers might interact with task situations in unexpected ways or things might go wrong both in terms of the user experience and in terms of the data quality. During these iterations the assessment designer often must go back and revise the theoretical student model because some variables of interest inevitably end up being more difficult to measure than initially anticipated.

The following sections of the chapter discuss the various phases of defining an assessment framework according to ECD. These phases are summarised in Figure 6.1. An *analysis of the target domain* – in other words, clearly defining its constituent knowledge, skills, attitudes and relevant contexts – lies at the foundation of any conceptual assessment framework. The following step is *domain modelling*, which involves translating these core definitions into assessment claims. This preparatory work on describing the domain orients the construction of more detailed test specifications through the three interconnected ECD models: the *student model*, the *task model* and the *evidence model*. Although the phases are described here sequentially, as earlier mentioned, in practice these phases are iterative.

Figure 6.1. Phases of defining the conceptual assessment framework in an ECD process



Establishing solid conceptual foundations: Domain analysis and modelling

Domain analysis

Domain analysis includes making an inventory of the concepts, language and tools that people use in the target domain, identifying the range of problems and situations in which people use those target knowledge and skills, and defining the characteristics of good performance in those domain contexts. There are many possible methods to gather this information. In traditional assessments of disciplinary subjects (e.g. mathematics), detailed descriptions of the domain are already available for use in assessment design. For international large-scale assessments, comparative reviews of curriculum content are regularly undertaken – for example for the International Association for the Evaluation of Educational Achievement (IEA) studies

(Twist and Fraillon, 2020[3]). However, this is generally not the case for assessments of complex competencies. This means that it is essential to rely on the contribution of a group of experts capable of constructing new representations of what expertise means in those domains using, to the extent possible, empirical observations.

Cognitive task analysis (CTA) is one useful method for identifying and understanding the behaviours that are associated with successful outcomes in complex problem situations (Clark et al., 2008[4]). CTA uses a variety of interview and observation strategies including process tracing to capture and describe how experts perform complex tasks. 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. These descriptions generated through CTA are then used to develop training experiences and assessments, as they allow assessment designers to identify salient features of tasks that are appropriate to include as well as identify decisions that are most indicative of expertise (see Chapter 4 of this report for an example of CTA used to define the assessment domain of complex problem solving in science and engineering as well as to inform subsequent assessment design via the lens of critical decision making).

The definition of an empirically-based model for 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 to achieve a performance goal and can produce exemplars of real work products or other tangible performance-based evidence that can be associated to proficiency claims.

Domain modelling

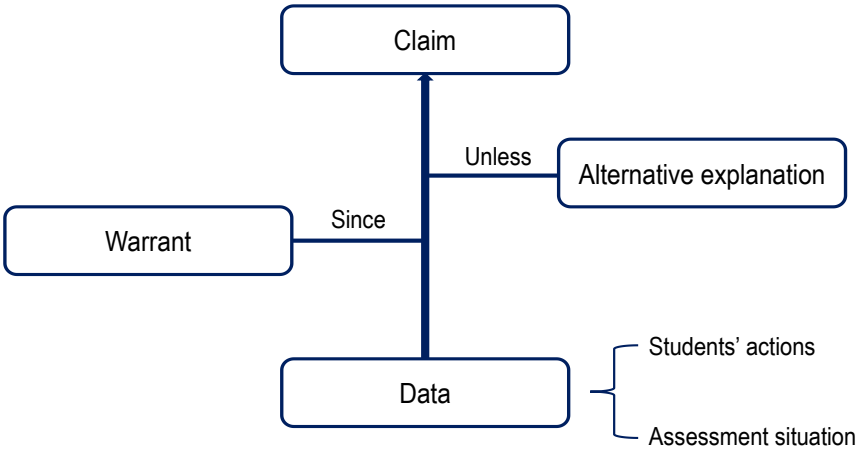
In the domain modelling phase, assessment designers collaborate with domain experts to organise the information collected during the domain analysis phase into assessment arguments (Figure 6.2). An assessment argument consists of three main elements: claims, data and warrants (Toulmin, 2003[5]; Mislevy and Riconscente, 2006[1]). A claim refers to what we wish to measure – for example, how proficiently a student solves mathematics problems or how capable they are of using language appropriately in an intercultural encounter. Data, such as the number of correct responses to test questions or behaviours observed in an interactive simulation, serve to support the claims. The warrant is the reasoning that explains why certain data should be considered appropriate evidence for certain claims.

Whenever an inference is based on complex data, such as those derived from actions in an open performance task, it becomes useful to add a fourth element to the assessment argument: alternative explanations. An alternative explanation refers to any other way(s) students could have done well in the test without engaging the relevant skills (e.g. gaming the system or guessing the right answer). For example, might a student struggle because of time pressure and not because they lack the relevant skills? Could the student be distracted by game-like aspects of the assessment?

Assessment arguments can be usefully formalised using "design patterns". Design patterns describe, in a narrative form, the student knowledge, skills and abilities (KSAs) that are the focus of the assessment, the potential observations, work products and rubrics that test designers may want to use, as well as characteristics and variable features of potential assessment tasks. The design pattern structure helps to identify which decisions have already been made and which still need to be made. Note that a design pattern does not include specific information about how materials will be presented to students in a given task nor about how scores will provide evidence about their proficiency in a domain; this more granular level of information is provided in the successive steps of ECD via the construction of the student, task and evidence models.

Figure 6.2. Assessment design as a process of argumentation

Toulmin's general schema for assessment arguments



Notes: Inference flows from data to claim by justification of a warrant; the inference may also need to be qualified by alternative explanations. Data are observed actions (by the student) that need to be contextualised in the specific assessment situation.
Source: Adapted from Mislevy and Riconscente (2006[1]).

The three models of an ECD assessment framework

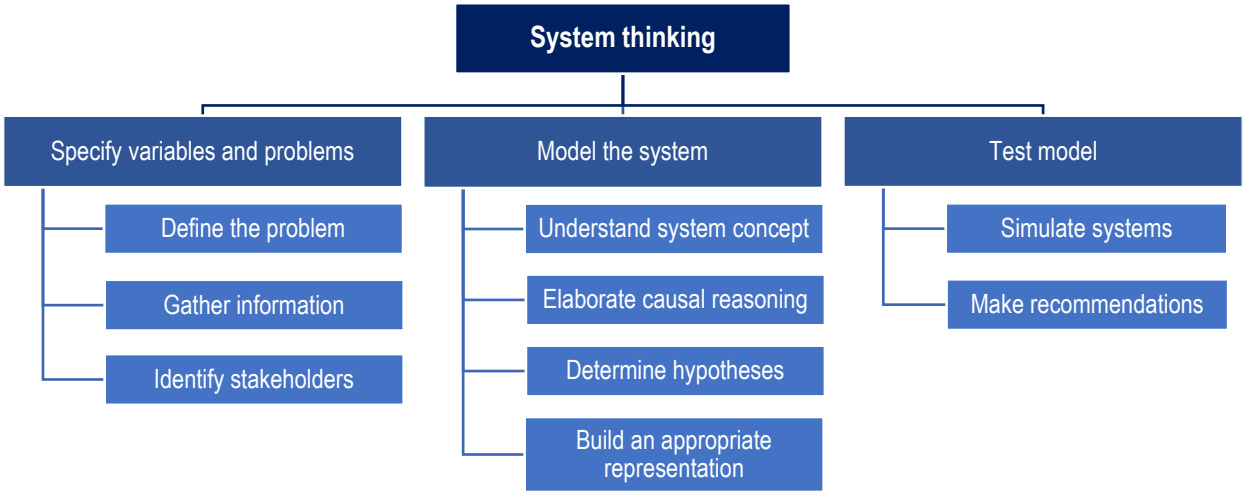
While the domain analysis and modelling phases provide the conceptual foundations for assessment, the three models of an ECD assessment framework guide the operationalisation of the assessment by providing substantive, technical and operational specifications used by task designers and data analysts.

The student model

The student model defines in detail the variables (KSAs) we expect to make claims about and the relationships between these variables. The student model is essentially the outcome of the assessment: it provides a map of each student's inferred KSAs, as specified in the domain model. In the simplest case the student model contains a single variable and student performance is computed from the proportion of tasks they answer correctly. For assessments of complex competencies with extended performance tasks, there are invariably multiple skills that, together, determine proficiency. In these assessments the variables in the student model generally include knowledge of key concepts in the domain, cognitive and metacognitive processes, mastery of specific work practices and strategies, and attitudinal, motivational and affect-regulation components.

The variables and their interconnections are often represented as a map in which each variable is a node connected by one or more links with other nodes. Figure 6.3 shows an example of a student model developed by Shute, Masduki and Donmez (2010[6]) for an assessment of system thinking using an immersive role-playing game. The model consists of three first-level variables: 1) specifying variables and problems in a system; 2) modelling the system; and 3) testing the model. Each of these first-level variables is further broken down into a number of second-level variables for which it is possible to define specific observables within the game environment corresponding to students' actions (e.g. interviewing stakeholders, collecting data, annotating hypotheses and conclusions, etc.).

Figure 6.3. An example of a student model for system thinking



Source: Adapted from Shute, Masduki and Donmez (2010[6]).

In the student model it is important to provide a vision for what the assessment can tell us about examinees' proficiency as they progress from the lowest to the highest levels of mastery in each of the constituent variables. Construct mapping is a design process that embeds this idea of progression, supporting the design of assessments that are closely connected to instructional goals. A construct map details what students understand and can do at incrementally higher levels of one or more proficiencies and indicates common misconceptions at each level (Wilson, 2009[7]). For example, in a construct map describing students' progression in their understanding of the Earth in the solar system, students at "level 3" proficiency think that it gets dark at night because the Earth goes around the Sun once a day (a common error for students at this level), while students a "level 4" no longer believe that the Earth orbits the Sun daily and instead understand that this occurs on an annual basis (Briggs et al., 2006[8]).

In most psychometric models the assumption is that the variables in the student model represent latent traits that are "fixed" at the time of taking the test. This assumption might not hold in innovative simulation-based assessments because students' knowledge or ability may change as they interact with the test environment. In other words, students may be able to learn during the assessment. This opportunity to learn during the test depends largely on the degree to which explicit and implicit feedback is integrated in the assessment tasks and environment (see Chapters 7 and 9 of this report for more on the design and implications of providing feedback as an intentional assessment design choice). Measuring the occurrence of learning within an assessment can help to identify those students who are more prepared for future learning. As discussed in Chapter 2 of this report, one core feature of next-generation assessments consists of shifting the focus of assessment from the reproduction of acquired knowledge to students' capacities to learn from new situations and using resources.

One issue this raises is how learning dynamics can be represented within a student model. Arieli-Attali and colleagues (2019[9]) have proposed to include additional "learning layers" for each of the three ECD models in order to describe the dynamics of change from one state of knowledge to another in digital environments. In their iteration of ECD, the "Student-Change model" specifies sequences of knowledge students are expected to acquire in the test environment and the processes they need to enact in order to move along this sequence. The sequence of knowledge components can be defined in terms of knowledge precursors needed to perform a given practice in the student model. For example, sentence comprehension relies on word identification, which in turn relies on letter recognition. The Knowledge-Learning-Instruction framework developed by Koedinger, Corbett and Perfetti (2012[10]) indicates three kinds of learning

processes: 1) memory and fluency building; 2) induction and refinement; and 3) understanding and sense-making. Specifying which kind of process is needed across all the learning sequences in the student-change model guides decisions about the scaffolding to include in the assessment environment. For example, if the focal learning process is fluency building, the environment should provide multiple opportunities to practice the relevant operation. In contrast if the focal learning process for a different KSA is understanding and sense-making, then the system should provide explanations and examples (Arieli-Attali et al., 2019[9]).

The task model

The task model defines the set of situations in which test takers can demonstrate their KSAs. It indicates appropriate stimulus materials, tasks and affordances, and clarifies variables that affect task difficulty as well as describes the work products that proficient students are expected to achieve (Mislevy et al., 2010[11]). The task model is closely connected to the student model as the tasks should be designed to elicit evidence for the KSAs. Multiple task models are possible in any given assessment, and different dimensions of proficiency might be best measured through different families of tasks.

Informative task models can be represented in different ways. One useful approach involves defining a Task-Model Grammar (TMG) following the principles of assessment engineering (Luecht, 2013[12]). The TMG approach provides an explicit description of: 1) the combination of KSAs needed to solve the task; 2) the types of declarative knowledge components that are typically used to challenge the examinee; 3) the complexity of the task components (e.g. investigating one simple variable vs. a complex system); 4) auxiliary information, resources or tools that should be embedded in the task; and 5) other relevant properties or attributes associated with each of the above components that might affect item difficulty.

There are no specific boundaries as to what constitutes an appropriate task. New ideas for engaging tasks are constantly emerging as information, interactivity and gaming technologies become increasingly integrated in assessment. Scalise and Gifford (2006[13]) presented a taxonomy of task formats in computer-based testing, largely distinguished by response type and level of interactivity. In this taxonomy, multiple-choice items represent the “most constrained” end of the taxonomy while presentations and portfolios represent the “least constrained” task formats; between them lies a range of other types of tasks including selection/identification, reordering/rearrangement, substitution/correction, completion and construction tasks. Chapter 7 of this report also presents a taxonomy of three broad task formats enabled in technology-enhanced assessments centred rather on the user experience and problem type presented to students. These three task formats include non-interactive problems, interactive problems with tools and immersive problem environments. Designers of complex assessments can draw on any of these technology-enabled task types when conceptualising their task models.

Unlike in standard assessments, evidence-bearing opportunities in complex assessment tasks are not limited to how students respond to pre-defined questions. Evidence of relevant KSAs can also be extracted by observing what test takers decide to do (or not) in a simulation or game – often referred to as “stealth assessment” (see Shute, Rahimi and Lu (2019[14]) for a review). For example, Wang, Shute and Moore (2015[15]) embedded an assessment of problem solving within the game Plants vs Zombies. The goal of the game was to protect a “home” base from zombie invasion by growing different types of powerful plants; to grow the plants, players needed to collect sun power by planting sunflowers. In this stealth assessment, players could demonstrate they “understand the givens and constraints in a problem” (one of the four variables in the student model) if they decided early in the game to plant sunflowers, as this action demonstrated that they understood that a lack of sun power was the primary constraint in the game. There was no explicit question in the assessment to elicit this behaviour; rather the relevant observables were obtained by tracking the sequence of test taker actions together with the corresponding changes in the game state. In games or simulations that are explicitly designed for assessment – see also the SimCityEdu example from Mislevy et al. (2014[16]) described in Box 3.1 (Chapter 3 of this report) – an essential part of

task modelling consists of defining the classes of actions that could be taken for each possible game state and associating these contextualised actions to unobservable KSAs (including no action taken at all).

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 real need to incorporate feedback or scaffolds in the task. However, as already discussed, innovative assessments might want to make claims about how students learn in authentic problem situations where they have access to resources and support. In the extended-ECD model developed by Arieli-Atteli and colleagues (2019[9]) and applied to the Holistic Educational Resources and Assessment (HERA) system for teaching and assessing scientific thinking skills, a new layer called the "Task-Support" model specifies the types of support that should be provided to students to facilitate their engagement in the learning progress(es) outlined in the "Student-Change" model. Three types of learning supports are offered after student errors: 1) "rephrase" (i.e. a rewording of the question that explains in more detail what is expected); 2) "break it down" (i.e. providing the first of several steps required to answer the question); and 3) "Teach me" (i.e. a written or visual explanation of the main concepts and operations required with illustrative examples).

Other ways to provide opportunities for learning exist that can be either on-demand or prompted by specific actions and outcomes in the environment (see, for example, Chapter 9 of this report). These resources must be designed with the same level of rigour that is put into the design of the task itself, establishing explicit connections to the claims one wants to make about students' use of these resources. For example, if the assessment aims to evaluate how students acquire fluency in a given skill (e.g. interpreting two-dimensional graphs), then providing them with a sequence of practice exercises focusing on that specific operation would be a justified resource. However, if the interest is rather on how students learn through transfer (i.e. applying concepts and skills to novel situations), then making worked examples available might represent a better way to provide scaffolding that is aligned with the assessment arguments.

The evidence model

Evidence models are the bridge between what students do in various situations (as described in task models) and what we want to infer about students' capabilities (as expressed in the student model variables). They specify how to assign values to observable variables and how to summarise the data into indicators or scales. The evidence model actually includes two interrelated components: evidence rules and the statistical model.

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. However, 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 student model (Mislevy, Steinberg and Almond, 2003[17]; Mislevy and Riconscente, 2006[1]). In simulation- or game-based assessments, evidence rules often rely on interpreting actions and behaviours that are recorded as process data (see Chapter 7 of this report for a description of different sources of process data). However, interpretation is susceptible to error as actions in 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 open and interactive 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 simulation (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.

In the process of defining evidence rules for complex assessments, it is fairly frequent that designers 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 mining.

Mislevy et al. (2012[18]) 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 that considered characteristics of test takers' submitted work products (for example, a given section of the network is considered "correct" if data successfully 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.

The second component of the evidence model 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 (e.g. responses, final work products, sequences of actions, etc.) and a student's 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 competence proficiency. More complex measurement models use latent variable frameworks, for example item response theory (IRT) (de Ayala, 2009[19]; Reckase, 2009[20]), diagnostic classification models (Rupp, Templin and Henson, 2010[21]) and Bayesian networks (Levy and Mislevy, 2004[22]; Conati, 2002[23]). A Bayesian network is a graphical representation of the conditional dependencies between observables and variables (nodes), and they constitute a particularly effective way of specifying, estimating and refining a measurement model in complex assessments. The conditional probability distributions of each variable are initially set following the opinions of domain experts and then refined empirically as data are accumulated. As indicators are produced by students (as defined in the evidence rules) their scores increase with respect to the relevant node of the student model, which then propagates through the network to increase the probability for the student's overall proficiency. Bayesian networks can also be combined with other analytical techniques (see Chapter 8 of this report for more on hybrid analytical models that combine strengths from different techniques).

Both the evidence rules and the statistical model can take into account the possibility of learning during a resource-rich assessment. If learning supports are available (e.g. hints, worked examples, scaffolds, etc.), students' decisions to use them can be modelled to make inferences about learning skills. Specific evidence rules are needed to specify how use of the learning supports can be interpreted as evidence of productive learning behaviour; these rules likely need to be conditional on the state of the test environment at the moment the student accesses the support. For example, the decision to consult a hint can be interpreted as evidence of productive help seeking *only if* the student has made (unsuccessful) attempts to solve the task on their own. In terms of the statistical model, changing levels of proficiency can be represented as a Hidden Markov model. The input-output structure of a Markov model allows for estimating the contribution of each support to a change in the latent KSAs, based on the change in the observed work product following the use of a support (Arieli-Attali et al., 2019[9]).

An illustrative example: The PISA 2025 Learning in the Digital World assessment

The previous sections of this chapter have underlined that assessment design is a complex process of decision making and iteration, guided by relevant theory, constraints, data collection and analysis. To

illustrate the complexities of this process using a tangible example of a complex assessment, this chapter provides a short overview of the decisions taken during the framework (and instrument) development of the PISA 2025 Learning in the Digital World (LDW) assessment. An interdisciplinary group of experts, coordinated by the OECD, led these decisions.

Domain analysis for LDW

Preparing the LDW assessment involved addressing a number of key questions, taking into account the specific constraints of the PISA test – namely: 1) What is the intended purpose of the test? 2) What type of learning is valued in the test? And 3) what concepts and practices are students expected to demonstrate and learn during the assessment? The purpose of the LDW assessment is to provide comparative, system-level data on students' preparedness to learn and solve open problems within digital learning environments. The focus of the assessment is motivated by the proliferation of digital learning resources and the rapid increase in their use as a result of the COVID-19 pandemic. However, while the potential of digital tools for empowering learning is broadly recognised in the literature and in the field, there is still insufficient evidence on whether students around the world are prepared to use these tools effectively.

Assessing how students engage in a learning process first requires identifying the type of learning we want to observe. The expert group decided to anchor the assessment in constructivist theories of learning. Constructivism is built on the belief that learners need to be active participants in the creation of their own knowledge and that students learn better if they possess a schema on which to build new understanding and link new concepts (National Research Council, 2000[24]). Social constructivists extended these ideas and emphasised the importance of interactions with other people and systems during learning. They contend that understanding is not constructed alone or within a vacuum but rather "co-constructed" through socially negotiated interactions with other people or objects. Assessments based on constructivism move away from focusing on discrete pieces of knowledge towards examining the more complex process of knowledge creation with external tools and resources, such as how individuals learn during effective, inquiry-based learning. For the LDW assessment, the expert group decided to focus on how well students can construct new knowledge and solutions through interactions with virtual tutors and digital tools.

Constructivism assigns a central importance to self-regulated learning and metacognition. Self-regulated learning (SRL) refers to the monitoring and control of one's metacognitive, cognitive, behavioural, motivational and affective processes while learning (Panadero, 2017[25]). Metacognition refers to the ability to monitor one's understanding and progress and to predict one's capacity to perform a given task (Connell, Campione and Brown, 1988[26]). Assessments that incorporate elements of metacognition ask students to reflect on what worked and what needs improving. From a design perspective, generating observations on SRL and metacognitive processes requires identifying problems that can only be solved through multiple iterations and developing test environments that include tools for monitoring and evaluating progress as well as providing feedback.

A key question during the domain analysis concerned what exactly students should learn and with what digital tools. In the digital world, tools abound that support the kinds of active and situated learning experiences that define constructivist approaches (see also Chapter 9 of this report for a more detailed description of the tools used to support self-regulated learning in the LDW assessment). Millions of students around the world design digital animations in Scratch (Maloney et al., 2010[27]) or explore scientific phenomena using PhET Interactive Simulations (Wieman, Adams and Perkins, 2008[28]). The experts referred to these practical examples from science, technology, engineering and mathematics (STEM) fields when imagining the type of work products that students should construct in the assessment as evidence of learning. Increasingly available evidence on how students' work with these tools also provided important empirical and theoretical references for defining the practices of proficient students and the typical struggles of beginners (see Brennan and Resnick (2012[29]) for an analysis of work practices in Scratch and de Jong, Sotiriou and Gillet (2014[30]) for an analysis of students' learning processes with scientific

simulations). The opportunity to rely on an established body of evidence on instructional design and learning progressions led the expert group to focus on the scientific practices of experimentation, modelling and design of algorithmic solutions as the "objects of learning" through which students could solve problems or understand complex phenomena. In those practices, it was deemed relatively easy to identify concepts and operations, such as the control of variable strategy or the use of conditional logic, that students could practice and "learn" in the short assessment time available.

Domain modelling for LDW

Domain modelling is essentially about defining assessment arguments. The LDW assessment aims to make multiple claims including: 1) Can students construct, refine and use models with the support of digital tools? 2) Can students define and apply algorithmic solutions to complex problems? 3) Do students seek feedback on their work or seek help when they are stuck? And 4) can students accurately evaluate their knowledge gaps and progress as they engage with complex challenges?

Let's consider the first of these claims in order to demonstrate how it can be represented through a design pattern for the LDW assessment (Table 6.1). As described earlier in this chapter, design patterns support test developers by providing a narrative description of the situations that the test should present to students as well as describing the core skills for which observables have to be generated, those that are not central to the claim but might affect performance, and the characteristic and variable features of possible tasks.

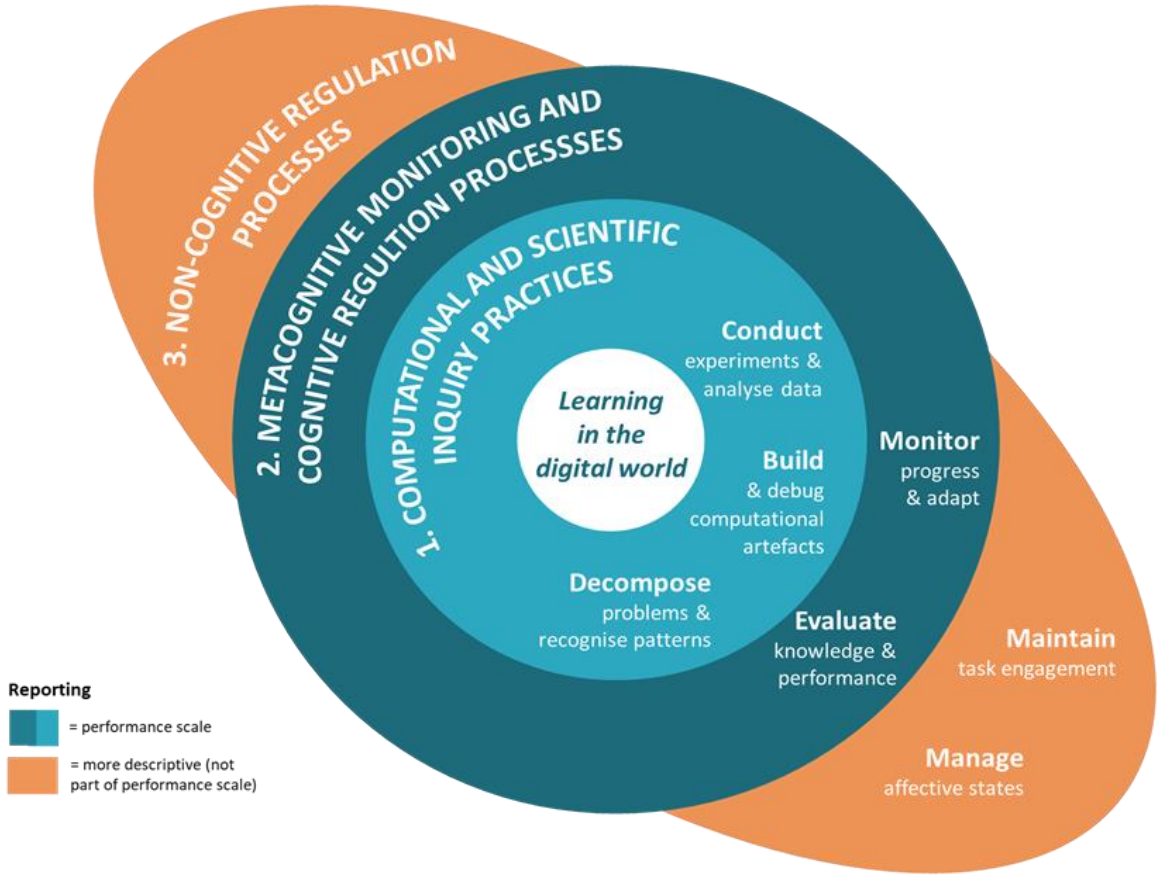
Table 6.1. Design patterns for the practice of modelling in the PISA 2025 LDW assessment

Rationale (warrant)	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.
Focal knowledge, skills, and attitudes	• Understanding the concept of variables, including dependent, independent, control and moderating variables • 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) • Identifying trends, anomalies or correlations in data • Experimenting using the control-of-variables strategy • Using a computational model to make predictions about the behaviour of a system
Additional knowledge, skills, and attitudes	• Functional knowledge of Information and Communication Technologies (ICT) • ICT self-efficacy • Prior knowledge of the phenomenon to be modelled • Perseverance, conscientiousness and mastery orientation
Potential observations and work products	• Student model represents the available information on the real-world situation • Student consults relevant information resources and collects relevant data to set the model parameters • Student modifies an incomplete or faulty model and justifies their modifications • Student identifies model weaknesses • Student uses their model to make correct predictions (given the available data)
Characteristic features of tasks	• Students are either provided with information about a real social or scientific phenomenon to model or provided with the tools to obtain this information • The student can check their model by comparing its output with real data • Students can use the model to make predictions
Variable features of tasks	• Level of familiarity of the phenomenon to model • Complexity of the ICT tools used for modelling • Student improves a basic model (provided to them) or builds the model from scratch • Student must find relevant data (in an information resource) or generate their own through experimentation • Number of variables to be modelled and structure of the system (simple vs. multi-level)
Constraints and challenges	• Limited time to learn how to use the modelling tool • Limited time to learn unfamiliar modelling concepts (e.g. control of variable strategy) • Large differences in prior knowledge in the target student population, meaning it is difficult to appropriately challenge all students on the same task

The student model in LDW

The domain analysis indicated that students can construct scientific knowledge and solve problems using digital tools if they master a set of computational thinking and scientific inquiry skills, can effectively monitor and evaluate their progress on a knowledge construction task, and can maintain motivation and manage their affective states. The student model for the LDW assessment is thus composed of three interconnected components: 1) computational and scientific inquiry practices; 2) metacognitive monitoring and cognitive regulation processes; and (3) non-cognitive regulation processes. Within each of the components, there are several facets (constituent skills).

Figure 6.4. Student model for the PISA 2025 LDW assessment



Source: OECD (forthcoming[31]).

Figure 6.4 illustrates the current version of the student model for the LDW assessment. At the core of the student model lies the construct, defined as "the capacity to engage in an iterative process of knowledge building and problem solving with digital tools" (OECD, forthcoming[31]). This definition of the construct with the notion of learning ("knowledge building and problem solving") at its core indicates that the assessment should not only provide measures of students' mastery of each of the constituent skills but also evaluate the extent to which students can progress towards their learning goals during the assessment.

Construct maps are currently being assembled for each of the facets of the student model (i.e. conduct experiments and analyse data, build and debug computational artefacts, etc.). The assessment development team is using data from observational studies and pilot tests to confirm initial hypotheses about what students are capable of doing at incremental levels of proficiency for each of the facets, and to

identify common errors and misconceptions at each level. Similarly, the team is working on defining the content and processes of learning that can occur during the assessment. This exercise is complex as the learning gradient depends to some extent on the initial level of knowledge of the student (i.e. students who start with a low level of initial knowledge might find it easier to progress using the learning resources in the assessment).

Task models for LDW

The development team identified two main typologies of units for this assessment. Units are sequences of tasks that are connected by the same scenario or learning goal. In the first type of unit, students construct an algorithm using different tools such as block-based programming or flow charts to solve a problem. In the second type, students represent and experiment with variables using a computational model that they can use to solve problems and make predictions (e.g. building an executable concept map where any change in one variable propagates to other linked variables). Drivers of complexity have been identified for both typologies of units (see Table 6.1). These relate to the accessibility of the digital tool that students are asked to use, the familiarity of the context of application, the intuitiveness of system dynamics and relationships between variables (in modelling tasks), and the complexity of the final product (in modelling tasks, the number of variables to experiment with or the presence of complex relationships such as moderating or random variables; or in programming tasks, the need to use complex control flows such as nested loops).

The team also worked on the definition of learning resources that could be standardised across test units. On one hand, these resources need to be useful so that students have an incentive to consult them; on the other hand, resources need to encourage transfer of concepts and should therefore avoid reducing the cognitive demand of the problem too much or directly giving students the solution. The experts decided to address this complex trade-off by choosing “worked examples” as the main learning resource. The worked examples provide an indication of how similar problems can be solved but students still have to transfer what they see in the example to a different context. Other opportunities for learning are provided by affordances to test programs and models or to check the correctness of one own’s solution after submission.

The aim of generating observables to evidence learning processes also demanded an innovative organisation of the unit. While discrete skills can be efficiently assessed through short and targeted questions, the process of knowledge construction can only be assessed through more extended tasks where students have opportunities to build on their initial understanding. Each unit has thus a duration of approximately thirty minutes and is organised as a series of phases. At the beginning, a virtual agent introduces the overall learning goal of the unit, framing the experience as a tutoring session (e.g. “I’m going to teach you how to…”). After this introduction, the students complete a pre-test that aims to measure students’ prior knowledge of the concepts and operations they will have to learn and apply in the unit. The third phase is a tutorial, where students are familiarised with the core functionalities and learning affordances of the digital learning interface. The fourth “learning phase” contains a series of discrete, carefully scaffolded and incrementally more complex tasks, the latter of which requires students to apply what they have learnt throughout the whole unit to a complex and multi-step problem. After the time available for learning and problem solving has expired, students complete some self-evaluation questions and report on their affective states during the unit. This complex design aims to immerse the students in an authentic digital experience where they are motivated to learn new things. Each unit contains a well-defined set of concepts and operations that students are expected to master following a coherent instructional sequence, starting from basic and progressing towards more complex applications. Two detailed prototype units are included in the framework to guide test developers in the application of these design decisions to the assessment tasks (OECD, forthcoming[31]).

Evidence model for LDW

The evidence needed to make claims on students' proficiency in the target KSAs is collected by checking the correctness of their responses to explicit questions in the learning and challenge phase, assessing the completeness of their work products (i.e. programs and models), and examining the strategies they followed as they worked as revealed by sequences of process data (log files). The LDW framework includes detailed evidence rules tables that describe how the observables above should be interpreted for scoring.

Some uses of process data for scoring are relatively straightforward. For example, in the modelling units, process data are used to check whether students completed all the steps required such as conducting a sufficient number of experiments to make an evidence-based conclusion on the relationship between two variables. Using process data is essential but much more complex for evaluating and interpreting SRL actions, given that any potential evidence of SRL behaviours must be evaluated in the context of the students' previous and following actions as well as the state of the environment when the action takes place. For example, the action of seeking help from the LDW tutor is considered evidence of proficiency in the facet "monitor progress and adapt" *only if* the student needs help (i.e. the student does not immediately seek help at the beginning of the task without first attempting to address the task) and *only if* the student does what the tutor recommends (where this is possible). The team has developed complex evidence rules for SRL actions that condition the scoring of potential evidence in this way. These evidence rules will also be complemented by applying data mining methods to pilot data to uncover other potential evidence of coherent SRL behaviours.

The statistical model used in this assessment will necessarily be more complex than those used in previous PISA tests. The main challenges in the LDW assessment lie in the local dependency of the measures (i.e. the tasks are not independent). Students' behaviours and actions in earlier tasks will likely influence subsequent performance. For example, those who carefully follow the tutorial and initial learning tasks are expected to perform better in the latter more challenging tasks. This violates the assumption of local independency of standard IRT models, requiring instead a model that is flexible and robust enough to account for dependencies across task observables.

Another complexity in defining the statistical model arises from the generation of non-random missing data. Some of the indicators for SRL are based on the choice of consulting learning resources; however, this choice is observed only for a fraction of self-selected students (i.e. those who struggle on the task are more likely to seek for resources). Tree-based item response models (IRTrees) represent one option that is being evaluated to address these issues of dependencies and non-random missing data (Jeon and De Boeck, 2015[32]). IRTrees models can deal with the dependence between observables that arise within extended tasks by capturing sequential processes as tree structures, where each branch finishes with a binary end-node. Alternative modelling approaches using dynamic Bayes nets are also being considered.

Conclusion

Thanks to progress in technology, it is now possible to build complex simulation environments that mirror or extend the real world and enable students to engage in the processes of making, communicating and interacting. Assessments using games, simulations or other digital media have the potential to generate better measures of deep learning, attitudinal beliefs and motivations, and thinking skills. In the context of international large-scale assessments, these technology-enabled innovations can serve to evaluate the efforts of education systems to support 21st Century competencies in more valid ways. However, designing new assessments in these dynamic environments is a complex exercise: designers must address multiple interconnected questions at the same time as confronting multiple uncertainties, ranging from the definition of the assessment arguments to the identification of statistical models that can deal with multidimensional and dynamic constructs.

Even more so than in assessments of established domains like reading and mathematics, the defensibility of any decision resulting from test scores crucially depends on early collaborative efforts to establish a solid assessment framework. As pointed out by Mislevy (2013[33]), close collaboration from the beginning of the design process is needed among users of the results (who understand the intended purposes of the assessment), domain experts (who know about the nature and progression in the target skills, the situations in which these skills are used and the behaviours that can be considered as evidence for these skills in these situations), psychometricians (who know about the requirements to achieve defensible reporting metrics), software designers (who build the infrastructure to bring the assessment to life) and user interface (UI) experts (who make sure that the assessment environment is intuitive to navigate). Not even the best and most complete team of designers can get everything right on the first iteration: by necessity, the design must be iterative and build on information collected through observational studies and small-scale validation efforts.

The design of the PISA LDW assessment has followed these principles and used the tools of ECD as described in this chapter. This was possible because more time and resources were available to develop the assessment framework and to undertake multiple design iterations following cognitive laboratories and pilot data collections than in previous PISA assessments. By articulating the complexity of design decisions behind an innovative test such as LDW, this chapter hopes to increase awareness about the fact that technology will deepen what we assess only if we multiply our efforts to learn how to use it within an evidence-centred framework.

References

Arieli-Attali, M. et al. (2019), "The expanded Evidence-Centered Design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design", *Frontiers in Psychology*, Vol. 10/853, pp. 1-17, https://doi.org/10.3389/fpsyg.2019.00853. [9]

Brennan, K. and M. Resnick (2012), "New frameworks for studying and assessing the development of computational thinking", *Proceedings of the 2012 Annual Meeting of the American Educational Research Association*, Vol. 1, http://scratched.gse.harvard.edu/ct/files/AERA2012.pdf (accessed on 27 March 2023). [29]

Briggs, D. et al. (2006), "Diagnostic assessment with ordered multiple-choice items", *Educational Assessment*, Vol. 11/1, pp. 33-63, https://doi.org/10.1207/s15326977ea1101_2. [8]

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. [4]

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. [23]

Connell, M., J. Campione and A. Brown (1988), "Metacognition: On the importance of understanding what you are doing", in Randall, C. and E. Silver (eds.), *Teaching and Assessing Mathematical Problem Solving*, National Council of Teachers of Mathematics, Reston. [26]

de Ayala, R. (2009), *The Theory and Practice of Item Response Theory*, Guilford Press. [19]

de Jong, T., S. Sotiriou and D. Gillet (2014), "Innovations in STEM education: the Go-Lab federation of online labs", *Smart Learning Environments*, Vol. 1/1, https://doi.org/10.1186/s40561-014-0003-6. [30]

Jeon, M. and P. De Boeck (2015), "A generalized item response tree model for psychological assessments", *Behavior Research Methods*, Vol. 48/3, pp. 1070-1085, https://doi.org/10.3758/s13428-015-0631-y. [32]

Koedinger, K., A. Corbett and C. Perfetti (2012), "The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning", *Cognitive Science*, Vol. 36/5, pp. 757-798, https://doi.org/10.1111/j.1551-6709.2012.01245.x. [10]

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. [22]

Luecht, R. (2013), "Assessment Engineering task model maps, task models and templates as a new way to develop and implement test specifications", *Journal of Applied Testing Technology*, Vol. 14/4, https://www.testpublishers.org/assets/documents/test%20specifications%20jatt%20special%20issue%2013.pdf. [12]

Maloney, J. et al. (2010), "The Scratch programming language and environment", *ACM Transactions on Computing Education*, Vol. 10/4, pp. 1-15, https://doi.org/10.1145/1868358.1868363. [27]

Messick, S. (1994), "Validity of pyschological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *ETS Research Report Series*, Vol. 1994/2, pp. i-28, https://doi.org/10.1002/j.2333-8504.1994.tb01618.x. [2]

Mislevy, R. (2013), "Evidence-centered design for simulation-based assessment", *Military Medicine*, Vol. 178/10S, pp. 107-114, https://doi.org/10.7205/milmed-d-13-00213. [33]

Mislevy, R. et al. (2010), "On the roles of external knowledge representations in assessment design", *Journal of Technology, Learning and Assessment*, Vol. 8/2, pp. 1-51, http://www.jtla.org. [11]

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. [18]

Mislevy, R. et al. (2014), *Psychometric Considerations in Game-Based Assessment*, GlassLab Research, Institute of Play, http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf (accessed on 27 March 2023). [16]

Mislevy, R. and M. Riconscente (2006), "Evidence-centered assessment design", in Downing, S. and T. Haladyna (eds.), *Handbook of Test Development*, Lawrence Erlbaum, Mahwah. [1]

Mislevy, R., L. Steinberg and R. Almond (2003), "Focus Article: On the structure of educational assessments", *Measurement: Interdisciplinary Research & Perspective*, Vol. 1/1, pp. 3-62, https://doi.org/10.1207/s15366359mea0101_02. [17]

National Research Council (2000), *How People Learn*, National Academies Press, Washington, D.C., https://doi.org/10.17226/9853. [24]

OECD (forthcoming), *The PISA 2025 Learning in the Digital World assessment framework (draft)*, OECD Publishing, Paris. [31]

Panadero, E. (2017), "A review of self-regulated learning: Six models and four directions for research", *Frontiers in Psychology*, Vol. 8/422, pp. 1-28, https://doi.org/10.3389/fpsyg.2017.00422. [25]

Reckase, M. (2009), *Multidimensional Item Response Theory*, Springer, New York, https://doi.org/10.1007/978-0-387-89976-3. [20]

Rupp, A., J. Templin and R. Henson (2010), *Diagnostic Measurement: Theory, Methods, and Applications*, Guilford Press, New York. [21]

Scalise, K. and B. Gifford (2006), "Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms", *Journal of Technology, Learning, and Assessment*, Vol. 4/6, http://www.jtla.org. [13]

Shute, V., I. Masduki and O. Donmez (2010), "Conceptual framework for modeling, assessing and supporting competencies within game environments", *Technology, Instruction, Cognition and Learning*, Vol. 8/2, pp. 137-161. [6]

Shute, V., S. Rahimi and X. Lu (2019), "Supporting learning in educational games: Promises and challenges", in Díaz, P. et al. (eds.), *Learning in a Digital World*, Smart Computing and Intelligence Series, Springer, Singapore, https://doi.org/10.1007/978-981-13-8265-9_4. [14]

Toulmin, S. (2003), *The Uses of Argument*, Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511840005. [5]

Twist, L. and J. Fraillon (2020), "Assessment content development", in Wagemaker, H. (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education Series, Springer, Cham, https://doi.org/10.1007/978-3-030-53081-5_4. [3]

Wang, L., V. Shute and G. Moore (2015), "Lessons learned and best practices of stealth assessment", *International Journal of Gaming and Computer-Mediated Simulations*, Vol. 7/4, pp. 66-87, https://doi.org/10.4018/ijgcms.2015100104. [15]

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. [28]

Wilson, M. (2009), "Measuring progressions: Assessment structures underlying a learning progression", *Journal of Research in Science Teaching*, Vol. 46/6, pp. 716-730, https://doi.org/10.1002/tea.20318. [7]

7 Designing innovative tasks and test environments

By John Sabatini and Xiangen Hu

(University of Memphis)

Mario Piacentini and Natalie Foster

(OECD)

Next-generation assessments of 21st Century competencies should confront students with relevant activities that are situated within authentic contexts of practice. This chapter unpacks the contemporary assessment designers' toolbox: it discusses how modern digital technologies can innovate assessment task formats, test features and sources of evidence to allow more interactive and immersive problems that are adaptive, include resources for learning and provide affordances for students to make choices. The chapter presents a framework that sets out technology-enhanced assessment task design possibilities and discusses the potential validity challenges and trade-offs that assessment developers will face when incorporating such innovations.

Introduction

Motivating the need for change in assessment design, How People Learn I and II (National Research Council, 1999[1]; National Academies of Science, Engineering and Medicine, 2018[2]) reviewed and described the multiple ways that individuals learn in distinct disciplines and domains on a trajectory towards expertise, mastery or proficiency. The result of successful learning is the ability to flexibly call upon one's knowledge and skills to identify and solve simple and complex problems in a domain – sometimes as an individual, sometimes collaboratively. As also argued elsewhere in this report (see Chapter 2), multiple voices in the field of assessment have applied this cognitive or learning science perspective to test design, proposing frameworks and models for transforming assessments from measures of static knowledge into measures with the twin purposes of evaluating an individual's position on a scale of expertise and drawing inferences about the kinds learning or instructional experiences that will likely advance them on this trajectory (Mislevy, 2018[3]; 2019[4]; Mislevy and Haertel, 2007[5]; Pellegrino, Chudowsky and Glaser, 2001[6]; Pellegrino, Baxter and Glaser, 1999[7]).

Assessing successful learning requires emulating the conditions in which knowledge and skills are applied. However, educational assessments have not always relied on the kinds of authentic tasks that enable one to evaluate the full range of constructs associated with a given competency – 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 – for example printing and transporting costs, test security, test environment, testing time and cost of scoring – while at the same time needing to satisfy technical psychometric standards of reliability, validity, comparability and fairness. Many of the features of "traditional" test design, administration, scoring and reporting (e.g. multiple-choice items) have taken shape because of such constraints (OECD, 2013[8]) and led to a predominantly paper-and-pencil testing regime administered in a single (often lengthy) session. Actual performance assessments were instead restricted to areas like the fine arts or spoken language and the need for one-on-one administration conditions and expert judge appraisal rendered them prohibitively expensive for use at scale. These various constraints have all contributed to delimiting the prospect of building the types of assessments called for by the learning sciences.

However, many of these 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 (see Chapter 5 of this report for an overview). In particular, the digital toolbox available to test developers now dramatically expands assessment design opportunities and affordances. For example, digital technologies create affordances for making the test experience less artificial (recall bubble-fill scantron answer sheets) and more face valid by approximating or simulating the situations or contexts in which target knowledge, skills, abilities and dispositions are used in real life. However, core steps in the assessment design process and validation of evidence remain conceptually the same and necessary to ensure that quality test results are produced.

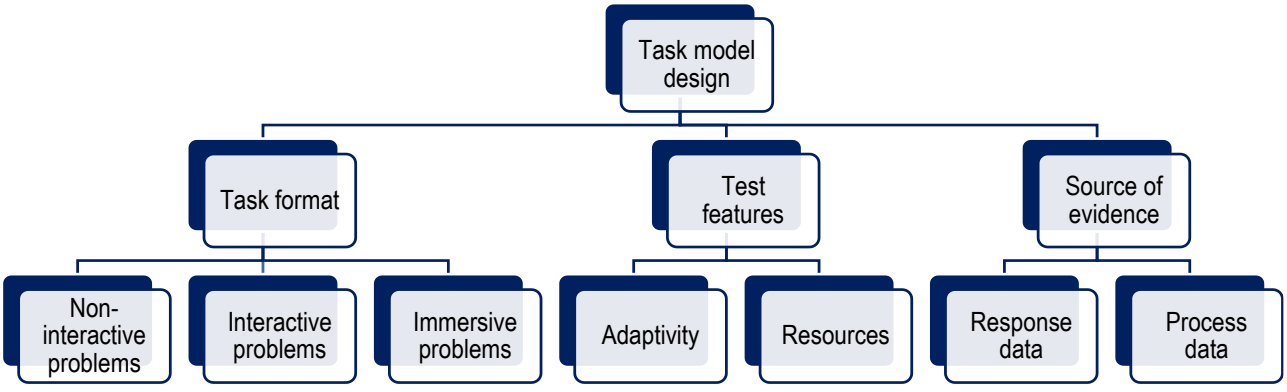
The preceding chapter (Chapter 6) detailed the core components of an Evidence-Centred Design (ECD) approach for making coherent and robust assessment design decisions, including defining the student model (i.e. the variables that we want to measure as they determine performance in a domain), the task model(s) (i.e. the situations that can elicit observations and potential evidence of such performance) and the evidence model (i.e. the scoring and accumulation of evidence to draw inferences about proficiency). In this chapter, we zoom into the task model component. We describe how technology-enabled innovations are or could be applied to enhance several aspects of task design – namely, task format, test features and sources of evidence – with the goal of eliciting and generating better evidence of what individuals know and can do.

Using technology to enhance task design

Assessment tasks should be designed to call upon the knowledge, skills, strategies and dispositions required to perform activities in the target domain, thus reflecting the student model. Key to this goal is what test takers encounter in a test: what they are expected to do, what they are able to do and the conditions in which they are expected to do it. Good assessment tasks should also be designed in anticipation of the subsequent interpretive and validation arguments both in terms of the information to be captured as potential evidence and how this evidence can be evaluated to produce valid inferences.

In the following sections, we propose a technology-enhanced assessment (TEA) task design framework that is organised around three main task model elements: 1) task format, or the kinds of problems that test takers engage with; 2) test features, or the affordances and features that can be enabled and overlaid with any of the task formats; and 3) sources of evidence, or the different kinds of observations that can be generated and captured as potential evidence. These elements are represented in Figure 7.1. We discuss the different possibilities enabled by technology in each of these three elements as well as the implications on validity of different design choices. In doing so we discuss a number of illustrative examples though by no means exhaustively cover the emerging possibilities from each cluster.

Figure 7.1. Technology-enhanced assessment (TEA) design framework



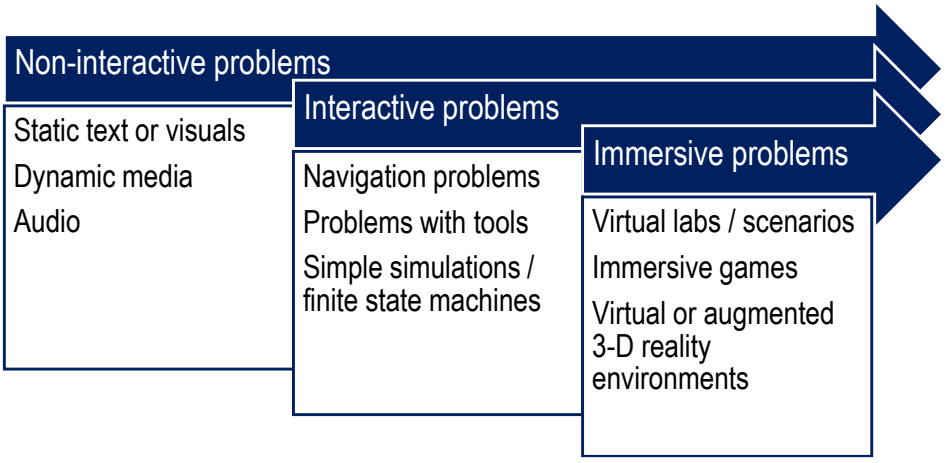
The discussion of the elements in the framework focuses on those possibilities now enhanced or enabled by TEA. That is not to say that all of the techniques cultivated in the paper-based testing era should be abandoned; rather, we contend that these types of items and test features are insufficient alone for measuring the types of complex competences advocated for in contemporary visions of education and assessment (including those in Part I of this report). The goal here is to highlight TEA task design possibilities for constructing more authentic tasks and producing potential evidence that can yield more valid inferences of what individuals know and can do, especially in more complex learning and problem-solving contexts. While several of the task design innovations we describe have been applied or researched to some extent in assessment design, to-date most have more extensively been researched in learning contexts as part of web- or software-based instructional designs.

Task format: Transitioning from static to dynamic to interactive to immersive

Technology-enhanced assessments significantly expand the types of task formats that can be presented to students. By "task format" we refer to the presentation of the main task content to test takers (i.e. the

task stimuli) as well as the nature of the task problem with which students are asked to engage. In this typology we acknowledge a continuum from static, non-interactive task formats to interactive problem-solving formats with (computational) tools to dynamic and immersive testing environments (see Figure 7.2). While we identify these three classes of task format made possible through TEA, we acknowledge that these are broad generalisations and that it is possible to combine features of the three classes within a single task – particularly when moving towards the more innovative end of the continuum (e.g. combining non-interactive stimuli like written text or audio recordings within immersive games).

Figure 7.2. Task format continuum in technology-enhanced assessment design



Non-interactive problems

Anchoring one end of the task format continuum are non-interactive problems. The types of fixed tasks used in traditional paper-based assessments presenting *static written text* or *visual stimuli* (e.g. photos, drawings, tables, maps, graphs or charts) to examinees can be found in this class of task format. Non-interactive task formats can also include more dynamic stimuli including *audio, animations, video,* and *other dynamic multimedia* content that are now relatively easy to produce in TEA. While both static and dynamic stimuli are possible in this task format class, the defining characteristics of non-interactive problems are that the 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 (beyond perhaps selecting amongst multiple-choice items or drag-and-drop functionality), and the test environment does not evolve as the test taker interacts with it.

Despite falling towards the “less innovative” end of the task format continuum, dynamic stimuli can nonetheless provide critical capabilities for enhancing the delivery of test content and therefore for enhancing validity. For example, well-designed animations or videos can increase engagement, help examinees to understand sequences of steps or actions when forming a mental model in domains such as science, and replace long written texts to reduce cognitive load and the dependency of test performance on reading skills. However, using richer media representations compounds the validity implications that should be considered. For example, there is a rich literature on the limitations and challenges students have in interpreting media, how/whether the media relate to textual descriptions, and generally skills in comprehension and interpretation of non-written stimuli (Mayer, Heiser and Lonn, 2001[9]).

Related, the TEA framework encompasses a vast range of assistive technologies including tools for magnification, guidelines for labelling visuals for text-to-speech capabilities, or the converse for speech-to-text functions. Deciding whether the examinee should have control over assistive technologies also has potential validity implications in terms of altering the cognitive demands of the task or creating unequal test

conditions depending on whether the examinee chooses to use them. For example, text-to-speech may benefit a struggling reader in a writing task but if the ability to read text is considered a pre-requisite proficiency to writing skills, then it can bias inferences of ability. Thus, while new task formats may provide new opportunities to enhance task presentation and accessibility, increase engagement and reduce cognitive load, these issues need to be investigated as part of any test validation processes.

Interactive problems

While non-interactive task formats may ensure a more uniform experience for test takers, these task formats can ultimately lack face validity especially when trying to replicate and assess behaviours in real-life problem situations. One major advantage of TEA is the ability to create or simulate interactive problem-solving scenarios that characterise more complex types of performances. These types of task format are more open and responsive to test taker actions and behaviours. They are typically multi-step, involve the use of *computer applications, tools* or *search engines,* and usually require *navigation* within and across screen displays.

The types of applications or tools now enabled in TEA environments are numerous and varied. As a simple example, consider word processing applications: they enable users to make notes and highlight text but can also assist with complex writing tasks beyond simply typing and formatting by correcting spelling, anticipating word selection, making recommendations to improve grammar, and providing tools like a thesaurus and dictionary. Similarly, spreadsheet or calculator applications aid in performing a variety of numerical and logical operations. Other more complex tools might include information repositories and search engines or those that enable modelling and simulations. These tools are increasingly used by students for solving problems in these domains; thus, it could be argued that they should become part of the assessment environment if we want to capture evidence that reflects real-life performance. In other words, although including such tools in assessment can scaffold the very skills we have traditionally assessed without their availability, it makes sense to incorporate them as natural supports in testing environments in the same way that pencil-and-paper maths tests allowed for the use of calculators.

In addition to providing tools and environments that better reflect contemporary contexts of practice, interactive problem task formats are necessary for assessing more complex performances. As argued in Chapter 1 of this report, complex constructs are defined at least in part by behaviours or processes meaning that tasks targeting their assessment need to generate evidence of how individuals behave in certain situations and iterate towards a solution (i.e. not only focus on eventual outcomes). Tasks therefore need to elicit target processes and behaviours in authentic ways – through open-ended and multi-step problems – and make those behaviours and thinking processes visible by providing students with tools to make choices and iterate upon their ideas. These task conditions cannot be provided by non-interactive task formats. In contrast, interactive problems allow students to engage actively in the processes of making and doing. For example, finite state machines (simple simulations involving fixed inputs, outputs and states) allow test takers to directly explore relationships and control systems, which can be contextualised in order to provide evidence about different kinds of problem solving (see the PISA 2012 creative problem solving assessment (OECD, 2017[10]) for examples in different contexts).

However, greater task openness and interactivity needs to be balanced with construct and practical considerations such as coverage of the domain or comparability. In other words, tasks that evolve as test takers interact with them may result in less uniform task experiences and therefore uneven coverage of the target constructs. This creates challenges in drawing inferences across student populations. Authentic and interactive tasks might also take more time to complete than simpler static tasks. Therefore, optimal task design for interactive problem types often involves optimising the trade-off between task authenticity and constraints, guided by the student model and practical considerations. Highly interactive and open tasks might adversely influence engagement by increasing the cognitive load for test takers especially if

not carefully scaffolded and designed in collaboration with user interface and user experience experts (Mayer, Heiser and Lonn, 2001[9]; Sireci and Zenisky, 2016[11]).

Interactive task formats will likely create new challenges for conventional approaches to establishing valid inferences from test scores. The inclusion of applications or tools might unfairly advantage some test takers over others based on individual differences like digital literacy or familiarity with certain software. Artificially created simulation environments will likely require some pre-training for test takers to fully understand how to operate them. Similarly, TEA tasks that require navigation and interaction inherently demand a certain pre-requisite level of familiarity with digital tools. These are all complex problems that will require new thinking about how to construct and gather evidence for interpretive and validity arguments. In general, decisions to include any tools and interactivity in tasks require test designers to examine possible sources of construct-irrelevant variance and consider how pre-requisite knowledge and skills for interacting with such TEA environments are reflected (or not) in the definition of the target construct.

Immersive problems

Anchoring the opposite end of the continuum (for now) from non-interactive problems are truly immersive problems. These include *simulated lab*, *immersive games*, or *3D modelling and virtual reality environments*. These types of immersive problems allow examinees to navigate through a two- or three-dimensional rendition of a virtual world, on a screen or via virtual reality (VR) headsets, which give the illusion of moving through space. The virtual world can be imaginary or real – one can interact with realistic simulated agents, such as teachers or peer students, or with fictional avatars or animals.

Immersive problems frequently employ game-based elements to enhance motivation as well as scaffold or control learner experience (Pellas et al., 2018[12]). One example of this paradigm is the virtual escape room, where an individual or group navigate a 3D-rendition of a space (room), solving large and small problems until they acquire the keys (i.e. meet criteria) to escape the room (Fotaris and Mastoras, 2019[13]). Shute (2011[14]) embedded an assessment of creative problem solving of this type in the 3D fantasy videogame Oblivion, where players have to solve quests such as locating a person or retrieving a magical object. Other examples of these types of assessment problems include simulations used most often for professional training, such as virtual aviation or medical intervention simulations. These types of tasks are becoming feasible to design, implement and scale, and one can imagine their increasing integration in more educational settings.

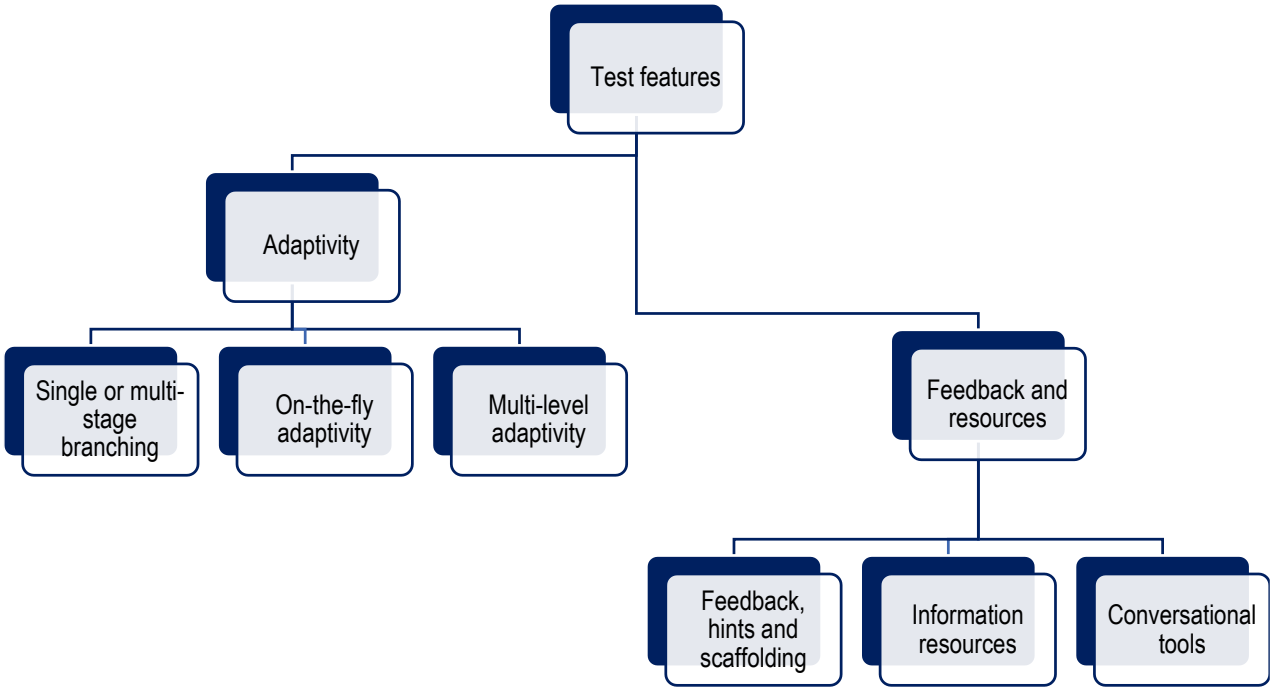
In addition to concerns about how to investigate validity as discussed for interactive tasks, research also does not yet show that immersive task formats always improve the measurement characteristics of a task. Indeed, Harris et al. (2020[15]) argue that virtual environments can often confuse the goals of presentation and function, relying on superficial visual features to “achieve” fidelity that are not the key determinants of validity. For example, face validity also requires that the structural and functional features of virtual simulations (e.g. how user inputs and behaviours relate to simulation outcomes) replicate those of real-life situations. It remains 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 and capture the use of knowledge and skills targeted in the student model (i.e. that they have construct validity), and that gamification or virtual immersion used to increase motivation and authenticity do not adversely distract from the task at hand (both literally for the student and figuratively speaking in terms of the assessment goal). Studies that demonstrate transfer from the test environment to real life applications of skills are as critical for establishing support for valid claims from immersive outcomes (i.e. predictive validity) as they are for traditional assessment designs.

Test features: Adaptivity and feedback

The second group of considerations relates to test features enabled by TEA. Here we discuss two specific features: 1) adaptivity; and 2) the provision of feedback and resources (Figure 7.3). Test features refer to

affordances or characteristics that can be overlaid with any of the task formats described in the previous section, in more or less transformative ways depending on explicit choices of assessment designers.

Figure 7.3. Test features in technology-enhanced assessment design



Adaptivity

Digital test delivery provides significant opportunities for integrating computer adaptive testing (CAT). In CAT designs, decision rules or algorithms select test items from an item pool for individual examinees. Techniques for implementing psychometrically viable adaptive designs have been established and refined for several decades; indeed, CAT (in its various formats) is one of the most researched innovations in test design since seminal work by Wainer and colleagues (2000[16]). CAT requirements often demand large item pools that are pre-calibrated using models such as item response theory (IRT) and the establishment of appropriate constraints. In this way, different examinees may take different items, but their scores are placed on a common scale. In large-scale testing, this type of adaptivity has been based largely on efficiently estimating the proficiency of an individual on a scale, and the selection of items is based on a set of constraints such as domain sampling and an estimate of ability calculated from performance on previous items.

Various CAT designs, with varying degrees of adaptivity, have been researched and implemented in large-scale assessment. Simpler adaptive designs utilise *single* or *multi-stage branching*: in these designs, test items are grouped into modules that differ in test item difficulty. There is typically one module in Stage 1, after which a computer algorithm selects one of the available modules in Stage 2 for examinees to sit next depending on their performance in Stage 1. Several designs are possible: there can be several modules within a stage, and multiple stages in a test (depending on module and test length); and different algorithms can be used to make branching decisions between stages. What these designs share is that adaptivity essentially occurs at the stage-level rather than the item-level. Other designs employ *on-the-fly adaptivity*, where adaptivity does occur at the item-level (i.e. each item is tailored to the student on the basis of 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. Such tasks may not be compatible with on-the-fly CAT designs. The strength of on-the-fly CAT approaches are that they are efficient in the delivery of items for the given constraint set within examinees' ability range. Consequently, such designs 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 as well as an arbitrary (rather than cohesive or thematic) trajectory through the content domain – thus highlighting the importance of the constraint set in any CAT design. Recent advances in CAT may help to address this issue by integrating hybrid measurement models – in other words, using techniques for modelling embedded construct dimensions such as multiple IRT, q-matrices, diagnostic modelling or knowledge tracing – which hold promise for measuring multiple aspects of performance and manipulating subsequent tasks attempted by the test taker. However, these designs are far less mature than their well-researched counterparts.

Another CAT option is to adapt tasks based on prior choices or actions of the examinee instead of an IRT-based parameter estimation algorithm. This is akin to the adaptivity typically found in videogames where the learner explores an open environment, and the environment adapts to their actions and behaviours within the environment. This design better reflects the reality of contingencies in problem solving environments and its strengths include greater authenticity, interactivity and adaptivity to learner responses – and if designed to give examinees some choice or control, it can also enhance engagement. However, enabling adaptivity fully on the basis of test taker choices can introduce substantial construct-irrelevant variance if choice is not included as part of the student model. Even in cases where choice is explicitly assessed, similar issues can arise as with on-the-fly adaptive models without sufficient constraint mechanisms. Another weakness of internally adaptive tasks is that they may 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 as well as more difficult to score for the purpose of standardised assessment. However, innovative assessments might be able to integrate this type of more complex and multi-level adaptivity by adopting some of the technical solutions already used in videogames that are designed to maintain player engagement by alternating states of learning and states of mastery (e.g. presenting a more challenging level or task whenever a simpler one is completed).

Feedback and resources

Feedback continues to be a frontier concept in assessment despite its prevalence and use in learning systems. Metacognition and self-regulation have become important concepts in psychology and the learning sciences. Metacognition refers to monitoring one's own cognitive processes and progress towards achieving a goal (Azevedo and Aleven, 2013[17]; Hacker, Dunlosky and Graesser, 2009[18]). Self-regulated learning refers to the regulation of one's cognitive and affective behaviours while executing a task including planning, setting goals, enacting and adapting strategies, and maintaining engagement and motivation (Panadero, 2017[19]). As argued in Chapter 2 of this report, iterative feedback cycles are central to learning and problem-solving processes. In the context of building adequate constructs of proficiency and therefore suitable next-generation assessments of those constructs, tasks need to be able to simulate these iterative feedback loops by creating opportunities for examinees to interact with feedback and resources.

One advantage of TEA is that test taker actions and responses can be immediately registered in the digital platform and algorithms can instantly score that data. The simplest form involves *solution feedback mechanisms* that inform test takers about the "correctness" of their response(s); these mechanisms could be triggered on-demand via a choice-based mechanism (i.e. explicitly sought by the test taker) or automated following a particular event (e.g. submitting a solution). More detailed feedback could also be provided, for example explaining why a given response is wrong or describing the optimal solution path. In such cases, especially when followed by allowing the respondent to try to answer again, feedback then functions as a type of *hint* or *scaffolding*. Hints and scaffolding aim to break tasks down into smaller more

manageable steps by giving pointed clues or guidance on what the test taker should do or by simplifying or partially completing the task at hand. In terms of the examinee experience, hint and scaffolding mechanisms can function similarly to solution feedback mechanisms via a choice-based mechanism (e.g. a hint button) or an automated action trigger (e.g. number of failed attempts, time delay, etc.).

Feedback, hints and scaffolds can be more or less "intelligent" in terms of the extent to which the support is tailored to a given student's needs. Fixed support (i.e. non-adaptive, where every examinee receives the same support regardless of their actions or solution quality) is clearly much simpler and more cost effective to develop and ensures a greater standardisation of test experience. However, the resulting trade-off is that such support is unlikely to be useful for all (or even the majority) of test takers; this may then be counterproductive, adversely impacting test taker engagement or affective regulation (e.g. triggering anxiety) that might jeopardise the validity of test scores. In closed tasks that employ selected response types (like multiple-choice items), "intelligent" feedback can be provided through relatively simple if-then branching based on students' responses. For example, different explanations might be provided to the test taker based on their selection of incorrect option (where each option reflects a particular or common misconception). However, for more open-ended tasks that involve interactivity and constructed responses, more complex algorithms are required that are capable of discerning solution quality and characteristics in addition to solution correctness. For example, more advanced analytical techniques may take into account sequences of prior responses or actions or may use natural language processing (NLP) to evaluate examinees' constructed responses, formulate targeted prompts/hints to the examinee, trigger scaffolding mechanisms or identify misconceptions (see Chapter 10 of this report for more on AI-enabled feedback).

Other learning supports can be provided to test takers through *information resources*. These might take the form of static written resources, such as a list of formulas or an encyclopaedia entry, or could be in any of the multimedia formats now supported by TEA (e.g. videos, animations, audio). More interactive information tools could also be provided to students, for example simulated or real search engines or interactive tutorials. In many cases, information resources function similarly to hints (when sought proactively by students); however, we consider hints as a source of more targeted, task-relevant information whereas information resources and tools may be less structured, more comprehensive and may not be task specific. While information resources are potentially less disruptive than other forms of interventionist scaffolding, hints and feedback, the research on providing optional information resources is mixed with respect to who uses which types and how effectively (Inglis et al., 2011[20]). Enabling choices in the context of providing learning supports in assessment integrates an additional construct in the student model. Choice therefore needs to be reflected in the understanding of the domain and incorporated into inferences being made about examinee performance. Perhaps one way to reason about whether to provide a particular resource is to ask whether a typical individual would have these kinds of resources available when solving this type of problem outside the test environment.

A third variation of feedback and information resources during assessment is to provide examinees with *conversational tools* through which they can engage with an informed or more-knowledgeable other. These tools can employ scripted or intelligent agents that can serve multiple functions including: 1) providing expert knowledge, instructions or guidelines; 2) modelling responses; 3) presenting alternative or contradictory points of view; or 4) reviewing or summarising content to help test takers' manage cognitive load. These agents may be personified via names and visuals to give examinees the illusion of interacting with specific individuals throughout the test experience and can be given specific roles within the task context (e.g. a peer student, teacher or content expert). One commonly used example in intelligent tutoring systems (ITS; see also Chapter 10 of this report) is a trialogue design in which a teacher agent and peer student agent interact with the test taker as they learn content and complete tasks, where each potential response to the examinee is scripted and a rule-based branching algorithm is used to decide on the most appropriate response the student receives (Graesser, Forsyth and Lehman, 2017[21]). In more complex ITS environments, agents can interact in more "intelligent" (i.e. non-scripted) ways with students through the AI-based analytical techniques described above. While scripted agents maximise standardisation and

scalability and minimise costs, intelligent agents can provide a more authentic experience by providing personalised feedback and enabling students to freely express their ideas (rather than selecting from pre-defined responses).

It is also possible to imagine cases where human agents (e.g. trained experts) or genuine student peers can communicate with one another multi-modally (via chat affordances or videoconferencing tools) or by working collaboratively on shared objects like a simulation; in these cases, genuine interactions can also constitute adaptive individual feedback or hints. However, in the case of peer-to-peer collaborative affordances, particularly in a large-scale context, the lack of standardisation over the feedback available constitutes a key challenge for understanding and making valid inferences from the situation (i.e. the quality of support available may depend on the proficiency level of one's peer).

With any of these feedback or resource affordances, decisions over the exact type and nature of support provided to students should ultimately be guided by the student model and the goal of assessment. For example, where the use of feedback is considered construct-relevant then assessment designers might consider it important to embed intelligent feedback mechanisms into tasks so that feedback is always useful to students. In other words, if all students receive the same feedback regardless of their solution quality at the time, then there will be some test takers for whom the feedback is not useful and who therefore cannot demonstrate the targeted skill. Similarly, where test taker choice is construct-relevant then perhaps an on-demand mechanism is appropriate; however, enabling choice may then preclude opportunities to observe such behaviours so it may be desirable to also build in some action- or event-triggered feedback mechanisms.

While any of the above elements may be warranted for inclusion in assessments tasks, especially if modelled in the domain, a further challenge comes in what scoring models to apply. As noted, research on these affordances has mostly taken place in the context of instruction. Feedback, learning supports and knowledge resources can potentially change the knowledge state of the examinee as the test proceeds therefore influencing an examinee's performance on future test items. In contemporary psychological models that incorporate metacognition and self-regulation, and in a world where feedback is a common and likely aspect of real world problem solving, it seems appropriate that authentic test environments incorporate feedback in task design. However, in traditional psychometric thought, these might be seen as violating item independence or interfering with the existing knowledge of the examinee; doing so therefore requires creating and validating new approaches to measurement models. While some progress has been made on this front (Levy, 2019[22]), 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, in feedback or hint models where the examinee is given more than one chance to respond, the scoring model might weight answering correctly the first time higher than requiring the hint or feedback; or 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 therefore critical for understanding how inferences can be made and what inferences can be supported as measurement claims when integrating such affordances in assessment.

Sources of evidence: Response product and response process data

The final group of elements in the framework relate to the expanded sources of (potential) evidence in TEA. The palette of potential evidence now goes well beyond traditional multiple choice or constructed (written) responses that have thus far dominated in large-scale assessment: data collected via TEA can include all nature of digitised logged responses and actions that can be generated in the task formats and interactive test design features previously discussed.

One distinction that has entered the conceptual research is between response product data versus response process data (see Table 7.1). Response processes "refer to the thought processes, strategies, approaches, and behaviours of examinees when they read, interpret and formulate solutions to

assessment tasks" (Ercikan and Pellegrino, 2017, p. 2[23]). Response processes also go beyond the cognitive realm, encompassing emotions, motivations and behaviours (Hubley and Zumbo, 2017[24]). Data that captures potential evidence of these response processes can therefore be understood as (response) process data; this 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 agents or human collaborators. Any data that might therefore be evaluated in understanding, characterising or evidencing student's thought processes, strategies, approaches and behaviours can be considered process data. Conversely, response products refer to students' final responses on an assessment task or a given item. Product data therefore typically refers to data resulting from selected responses (e.g. multiple-choice items), short or extended written responses, or the final product in a simulated or performance demonstration.

Table 7.1. Sources of evidence in technology-enhanced assessment

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, etc.)
Written response (short or extended)	Intermediate solution states (i.e. those before submitting final solution)
Spoken response	Action logs (e.g. use of affordances, keystrokes, mouse clicks, events, etc.)
Performance response (e.g. level attainment in a game, simulation state, artefact)	Physiological measures (e.g. eye-tracking data)

Product data

The simplest form of product data is generated through selected response formats. Selected responses are typified by *multiple-choice* or *true/false items* that typically present students with pre-defined answer options (for which there is one optimal solution). Often the choices presented to students include "distractors" that might appear plausible to (some) students but should not be justifiable as a correct response. When feasible, distractors should be designed to reflect understanding of students' mental models in the content area such that students' choice of distractors can also provide useful information about their proficiency level. Integrating some form of adaptive branching can also enable more complex selected responses whose answer options change on the basis of students' previous selections (e.g. asking students to choose an explanation for their previous answer).

TEA designs also have the potential to make selected response type formats more engaging by re-imagining visual formats and introducing elements of test taker interactivity, for example *drag-and-drop* options for filling in tables or graphic arrays, identifying locations on the screen via visual mapping (*hotspots*), or *highlighting or selecting text or objects* in an array. These more interactive variations allow some degree of active response (i.e. actively rearranging elements into the correct order rather than choosing from a pre-defined list). Selected responses can also be combined in stealth ways with more interactive problem types. For example, as part of a simulation-based task a student might be required to set the parameters for conducting an experiment from a range of possible choices (i.e. they can choose all that apply in that situation). While in a real world situation there might be one "best" combination of choices, there are also likely other combinations that are partially correct. In these types of more complex and open task situations it makes sense to use a scoring rubric that rewards different combinations of selected response items with different scores.

Despite being more commonly used to produce scores than other sources of potential evidence – both due to easier interpretation and potential for automation – data generated through selected response types have some significant weaknesses in terms of validity. Because answers are (at least to some extent) pre-

defined, test takers may guess the correct answer. These response types also cannot provide direct evidence of production skills. In contrast, constructed response formats necessarily require students to engage in some kind of a production activity and consequently are far less susceptible to unduly rewarding or eliciting guessing behaviours. However, constructed response formats and performance tasks require greater engagement and motivation on the part of test takers as well as generating richer data and potentially less-structured data that can be more complex to score in a reliable and comparable way. A further limitation of constructed responses is imposed by their scoring models which may take the form of rubrics or guidelines. These often restrict task design to elicit the types of responses that can be scored by rubrics or trained scorers, thus reducing the complexity of authentic tasks in order to obtain reliable and consistent expert judgments of quality.

A common constructed response type is *written responses*, which can range from short written words or sentences to extended essays. Advances in technology and data analytics now mean that it is increasingly possible to score written responses using techniques such as NLP coupled with machine learning, therefore reducing the burden on human scoring. For example, if a given language and region already has a community that offers software or a database for linguistics analysis (e.g. the Linguistic Data Consortium (n.d.[25]) for English, United States), syntactic analytical tools can be used to evaluate the grammaticality of short answers. In the absence of large datasets, machine learning algorithms can also be trained to identify key criteria such as semantic similarity with the answer keys.

Both written and *spoken responses* have typically been scored against rubrics involving expert human judges. Technological advances including improving the quality and automated transcription of audio recordings, increases in bandwidth and digital storage, and speech recognition software capabilities are converging to remove the barriers for this type of response option to be used in large-scale assessment. This an area of exponential growth stemming from advances in Artificial Intelligence (AI) but scoring any assessments – particularly those with stakes attached – using such techniques is still a work in progress.

Often the users of Machine Learning (ML) or AI-derived test scores remain wary or suspicious of machine scoring approaches, creating an extra burden on test producers to communicate results to users. It is best that any consideration of an AI-based scoring engines should be taken into account during the design stage, and that results from such assessments be properly scrutinised in terms of threats to validity of score inferences such as cross-cultural comparability (see Chapter 11 of this report for an in-depth discussion of this and other issues on cross-cultural validity of test scores from innovative assessments).

As in any assessment design process, a main consideration for test designers is whether a chosen response type (and therefore source of product data) is a valid and appropriate modality for expressing the answer to a particular task or question; and as a corollary, whether that response type would disadvantage any sub-population in comparison to alternative modalities. These considerations are not new and should be applied in the context of all choices regarding response types including the use of more “traditional” formats like multiple-choice or short written responses. What is new, however, are potential alternatives to such formats that may elicit evidence of similar aspects of constructs in different ways.

Finally, TEA enable other types of *performance responses* that can generate product data and constitute potential evidence of students’ proficiency in learning and problem solving. Interactive problems with tools enable students to produce complex but tangible artefacts or products. For example in specialised fields like architecture, exams have included graphical software tools that are common to the profession so that an examinee can create realistic designs in response to authentic problems. This idea can be generalised to any kind of visual, graphic, audio or computational (e.g. programming, modelling) tools available in a digital format that test takers can use to produce a tangible product for scoring. Other performance responses might include instances where test taker behaviours within an interactive or immersive problem bring about changes in the test environment (e.g. completing a level in a game or changing the state of a simulation). A key validity issue with respect to performance responses becomes one of examinee prior experience and training using the digital tools and environments provided. If one can expect that

examinees should have relevant experience or that the use of tools is an inherent part of the construct (i.e. it is construct-relevant), then it is only a matter of including the appropriate tool sets. However, if the tools serve to create an environment in which performance can be demonstrated but they are not inherently a part of the target construct, then familiarity with the tool/environment is not a pre-requisite skill and appropriate tutoring or training may be necessary to ensure fairness and valid results.

Process data

One of the most significant breakthroughs of the digital age in the context of assessment is the ability to control and capture fine-grain, real time process data as the examinee completes an assessment. Process data can be exceptionally varied and the same source of data can provide potential evidence for different aspects of validity and performance. As explained in greater detail in Chapter 12 of this report, there are two major complementary challenges when using process data in assessment. First is the matter of interpretation: the sheer report and complexity of the data produced and captured by TEA, not to mention its relationship to the construct, makes valid interpretations of test takers' behaviours and thinking processes a key concern. For example, *action log data* can provide information related to content (i.e. which key or affordance is selected), time (start, stop duration) and sequence of actions that may or may not reflect important information about the construct being examined.

Large-scale data analytics may be helpful with the treatment of process data but that leads to the second challenge – the matter of purpose. Process data is most likely to be of value in the design process when generated through intentional task design and in the context of understanding expected responses. Ideally, task designers would plan ahead for applying process data analytics to specific theoretically-derived hypotheses about expected patterns or strategies used by examinees approaching a task. For example, do experts take different strategic pathways from weaker examinees or novices as expected?

Emphasising the importance of theory-driven analysis is not to say that more exploratory analysis is unimportant; practically speaking, expecting a task designer to imagine all potentially valuable process data patterns *a priori* seems unfair and perhaps unwise. Here, exploratory research can reveal new insights, both practical and theoretical, that can feed forward into better and more efficient task designs that capture relevant interpretable data for use in evidence models. For example, are there multiple pathways to correct answers that were not originally envisaged, including some that bypass use of the target skill or construct that the task intended to measure? How do emotional responses (anxiety, boredom), engagement and motivational factors interact with performance? In closing, we note that the distinction between product and process data is relative to purpose – in other words, how the data will be used in the validation process – which again calls for close consideration of what counts as construct-relevant versus irrelevant.

Conclusion

In this chapter we have presented a framework for how technology can become an integral part of assessment task design. We highlighted how technologies can be applied when creating tasks, capturing new aspects of open and interactive performance environments where students can explore resources, engage in iterative processes of problem solving, and get real-time feedback on their progress. The TEA framework describes the expanded toolkit for assessment designers that stem from technological advances in the field. We organised the framework around three interrelated design issues that assessment designers must simultaneously consider: 1) task problem format (i.e. the kinds of problems and the level of interactivity that tasks present to students); 2) test features (i.e. the affordances and level of adaptivity that the test environment provides); and 3) sources of evidence (i.e. the classes of responses and the types of data that are elicited and used for scoring and interpretation). These design issues represent both a complex and simple challenge to the test designer: they are complex in that there are

now many more design possibilities and tools to consider when developing coherent tasks for a given test form; yet the challenge is also simpler, in that the designer can now draw inspiration from how tasks are performed in the real world, replicating them in a measurement environment with a broad palette of technological tools at their disposal to help them enact their vision of task (and evidence) models.

References

Azevedo, R. and V. Aleven (2013), "Metacognition and learning technologies: An overview of current interdisciplinary research", in *International Handbook of Metacognition and Learning Technologies*, Springer, New York, https://doi.org/10.1007/978-1-4419-5546-3_1. [17]

Ercikan, K. and J. Pellegrino (eds.) (2017), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, https://doi.org/10.4324/9781315708591. [23]

Fotaris, P. and T. Mastoras (2019), "Escape rooms for learning: A systematic review", *Proceedings of the 13th International Conference on Game Based Learning*, https://doi.org/10.34190/GBL.19.179. [13]

Graesser, A., C. Forsyth and B. Lehman (2017), "Two heads may be better than one: Learning from computer agents in conversational trialogues", *Teachers College Record: The Voice of Scholarship in Education*, Vol. 119/3, pp. 1-20, https://doi.org/10.1177/016146811711900309. [21]

Hacker, D., J. Dunlosky and A. Graesser (eds.) (2009), *Handbook of Metacognition in Education*, Routledge, New York, https://doi.org/10.4324/9780203876428. [18]

Harris, D. et al. (2020), "A framework for the testing and validation of simulated environments in experimentation and training", *Frontiers in Psychology*, Vol. 11, https://doi.org/10.3389/fpsyg.2020.00605. [15]

Hubley, A. and B. Zumbo (2017), "Response processes in the context of validity: Setting the stage", in Zumbo, B. and A. Hubley (eds.), *Understanding and Investigating Response Processes in Validation Research*, Social Indicators Research Series, Springer, Cham, https://doi.org/10.1007/978-3-319-56129-5_1. [24]

Inglis, M. et al. (2011), "Individual differences in students' use of optional learning resources", *Journal of Computer Assisted Learning*, Vol. 27/6, pp. 490-502, https://doi.org/10.1111/j.1365-2729.2011.00417.x. [20]

Levy, R. (2019), "Dynamic Bayesian Network modeling of game-based diagnostic assessments", *Multivariate Behavioral Research*, Vol. 54/6, pp. 771-794, https://doi.org/10.1080/00273171.2019.1590794. [22]

Linguistic Data Consortium (n.d.), *www.ldc.upenn.edu*. [25]

Mayer, R., J. Heiser and S. Lonn (2001), "Cognitive constraints on multimedia learning: When presenting more material results in less understanding.", *Journal of Educational Psychology*, Vol. 93/1, pp. 187-198, https://doi.org/10.1037/0022-0663.93.1.187. [9]

Mislevy, R. (2019), "Advances in measurement and cognition", *The ANNALS of the American Academy of Political and Social Science*, Vol. 683/1, pp. 164-182, https://doi.org/10.1177/0002716219843816. [4]

Mislevy, R. (2018), *Sociocognitive Foundations of Educational Measurement*, Routledge, New York, https://doi.org/10.4324/9781315871691. [3]

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. [5]

National Academies of Science, Engineering and Medicine (2018), *How People Learn II: Learners, Contexts and Cultures*, National Academies Press, Washington, D.C., https://doi.org/10.17226/24783. [2]

National Research Council (1999), *How People Learn: Brain, Mind, Experience and School*, National Academies Press, Washington, D.C., https://doi.org/10.17226/9853. [1]

OECD (2017), "PISA 2015 collaborative problem-solving framework", in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*, OECD Publishing, Paris, https://doi.org/10.1787/9789264281820-8-en. [10]

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=18A9CC493392BE9A918508D9929D29A3. [8]

Panadero, E. (2017), "A review of self-regulated learning: Six models and four directions for research", *Frontiers in Psychology*, Vol. 8/422, pp. 1-28, https://doi.org/10.3389/fpsyg.2017.00422. [19]

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. [12]

Pellegrino, J., G. Baxter and R. Glaser (1999), "Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice", *Review of Research in Education*, Vol. 24/1, pp. 307-353, https://doi.org/10.3102/0091732x024001307. [7]

Pellegrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know: The Science and Design of Educational Assessment*, National Academies Press, Washington, D.C. [6]

Shute, V. (2011), "Stealth assessment in computer-based games to support learning", in Tobias, S. and J. Fletcher (eds.), *Computer Games and Instruction*, Information Age Publishing. [14]

Sireci, S. and A. Zenisky (2016), "Computerized innovative item formats: Achievement and credentialing", in Lane, S., M. Raymond and T. Haladyna (eds.), *Handbook of Test Development*, Routledge, New York. [11]

Wainer, H. et al. (2000), *Computerized Adaptive Testing*, Routledge, New York, https://doi.org/10.4324/9781410605931. [16]

8 Analysing and integrating new sources of data reliably in innovative assessments

By Kathleen Scalise, Cassandra Malcom and Errol Kaylor

(University of Oregon)

This chapter explores whether robust analytic techniques are available to generate defensible inferences from complex data generated by digital assessments, including process data. Digital technologies hold great promise for helping to bring about changes in educational measurement and assessment, but one challenge to be faced is how to accumulate different sources of evidence in defensible ways to make inferences when constructs and observations include inherent complexity. This chapter discusses some potential solutions for making measurement inferences at scale using complex data, notably hybrid measurement models that incorporate one measurement model within another. The chapter conceptually draws on an example of a complex task in a technology-rich environment from the OECD Platform for Innovative Learning Assessments (PILA) to highlight this analytical approach.

Introduction

This chapter addresses some analytical approaches for including complex process and interaction data from technology-based tasks into educational assessments. An intersection is emerging between applications of learning analytics and traditional psychometric approaches to educational measurement (Papamitsiou and Economides, 2014[1]; Scalise, Wilson and Gochyyev, 2021[2]). This emerging intersection may help assessments to tap new technology for analysing and integrating different sources of data reliably in innovative assessments. By establishing some common ground between fields, new approaches called "hybridised" may borrow strength across toolkits to good effect.

Do we need new analytic approaches?

When tackling the topic of next-generation analytics for next-generation educational assessment tasks, the first question one might ask is whether new analytical approaches are needed. This is a typical "double-barrelled" question – or even "triple-barrelled." That is, there are at least three questions involved here: first, is there value in leveraging more complex and richer data in educational assessments available through new technology tools? Secondly, do we still need traditional response data? Finally, are sufficient techniques already available to accumulate the evidence from such tasks and make robust and interpretable inferences? The general consensus of this report regarding the first two questions is "yes", based on the earlier chapters. We will not explore the first two questions further because they are addressed elsewhere in this report (see Chapters 5, 7 and 12, particularly). Instead, we turn to the third question: Are robust techniques available to generate defensible inferences from such complex content?

Part of the answer depends on what is meant by "defensible" inferences. Over six decades of research in psychometrics and measurement technology for summative assessment has matured to establish well-accepted procedures for important issues, 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. Much of this is done using well-fitting measurement models. According to Levy (2012[3]) and others, new approaches using more complex and noisier data such as process data are still in their infancy and cannot rely on such procedures to establish validity if they are not well-fitting to measurement models. We will call this the "Levy challenge".

Chapter 5 of this report argues that information technology provides new opportunities to assess some hard-to-measure 21st Century competencies. Chapter 6 discusses the decisions involved in translating a theoretical model of a complex domain into an operational assessment, including when the domain involves observing dynamic states over time and capturing processes via technology-based simulations. Chapter 7 provides an overview of how information technologies can support scaffolding, feedback and choice while presenting video, audio, agent-based and immersive experiences. All of these information technology advances are very compelling, but a key question remains: has *measurement* technology evolved far enough to tell us how all these new *information* technology affordances can generate defensible measurement claims and allow us to make interpretable inferences about respondents or the groups they represent?

While many innovative psychometric models exist that can handle a variety of complexities in measurement, these are often employed in research projects only. Conversely, psychometric models used operationally do not well incorporate complex data (which can include but not be limited to process data). For instance, achieving goals in a serious gaming product or simulation setting may involve complex data such as examining an outcome figure with computer vision. While machine learning (ML)/Artificial Intelligence (AI) engines along with innovative psychometric models have been proposed in research to handle a variety of situations, operational models tend to better handle single dimensions and standardised

cross-sectional data better than complex data. Operational models tend to require very robust characteristics at the item level (that is, for a single question or a single observation) and do not work for assessments where different tasks are connected, and examinees' behaviours can only be interpreted by considering patterns of combined observations. These traditional models therefore often narrow constructs, with those aspects of constructs that are not yet well understood arguably the most vulnerable to inadvertent pruning. Scores or categorisations that are intended to provide evidence for different facets of a construct often are "collapsed" after delivery of an innovative task by combining score categories, with the consequence of obscuring different strategies for reaching the score – highly instructionally-relevant information for teachers and other users of the assessment results. Highly constrained assessments can generate relatively sparse, clean and standardised data for which such models are designed, but these assessments leave limited room for agency, adaptivity, cultural relevance, maker culture, collaboration or other potentially desirable aspects of student performance.

Some possible solutions

Such issues related to accumulating complex bits of evidence to make valid and defensible inferences suggest that existing analytic approaches are unlikely to be the next-generation models needed for large-scale summative assessments without some degree of revision and update. Although we argue elsewhere (see Chapter 13 of this report) that much work remains to bridge the broader fields of learning analytics and psychometrics to develop innovative measurement solutions, we summarise here a few directions of possible solutions for the technology-enhanced realm:

- *Borrowing strength from each other directly within the analytic techniques,* for example incorporating one measurement model within another or using one model to extend another. We describe this approach in more detail later in the chapter.
- *Establishing multiple inferential grain sizes*, for example designing a task to report out at different grain sizes for different purposes. The results from different analyses should be triangulated to ensure that there is consistency in the overall communication about students' proficiency.
- *Drawing on stronger confirmatory data from improved domain modelling and analysis*, for example drawing on strong theoretical research about learning patterns to design a task and then collecting data, analysing and then revising tasks and constructs iteratively based on findings as more data are gathered (see also Chapter 6 of this report). This approach contrasts with "dropping" more complex tasks that don't fit as well as more traditional ones after field trial testing, for instance – although it does require more flexible budgets and timelines for test development.
- *Using a richer range of exploratory data to explain more variance*, which can be done in a variety of ways such as including process data. However, such results may originate a theory through data mining rather than confirming one with a designed experience based on research, such as described in the first and third bullet points. If data mining is intended to generate a metric, it implies a couple of different commitments. Initial results – even if at scale – need to be interpreted cautiously, even though this flies in the face of making large claims to gain funding and support for projects. This therefore interacts with assessment literacy about the use of evidence (i.e. will the information, once available, be employed cautiously?). It also implies a commitment to investing considerable resources over time, ultimately converting from an exploratory to a confirmatory position with additional data sets and a clear interpretation.
- *Functioning with much more (and noisier) data* by attempting to use different types of technology-enhanced tasks to measure the same construct. This may reduce validity issues related to results being dependent on the technology presented in the task; however, it becomes an empirical question of whether a generalisable trait is being tapped across tasks. Technology-based tasks can generate a big data stream, but it is often unclear in what ways the data are "contaminated"

with construct-irrelevant variance. Typically, this is resolved by making inferences only over a body of different tasks, observations or questions rather than depending too much on one type – but this often implies a willingness to accept somewhat noisier tasks (or in formal measurement language, noisier testlets). Even when very carefully designed, individual tasks or testlets in such assessments tend to pick up on at least some constructs not being measured, including format effects, the context, passage, phenomena or other non-standardised elements. Trying to discard unique variance from technology tasks to solve this in a simple statistical adjustment, such as with bivariate models or even testlet models, also can discard salient variance unique to that task. To avoid this, the assessment blueprint needs duplicate tasks on the same standards, blocks of individual independent questions and good attention to testlet effect reduction. These requirements may mean time is too restricted to meet the evidentiary need if only large-scale assessments with relatively short testing times are used for inferences, as compared to inferences from well-developed systems of assessment such as multiple assessments designed to aggregate effectively together that are less constrained by time.

Understanding innovation challenges from a measurement perspective

To find a measurement solution for innovative tasks that generate complex data, the first step is to broadly understand the measurement challenges your types of innovation will face. To discover this, the key elements of an innovative task can be mapped into a conceptual space that can help describe it for measurement purposes. A taxonomy of five conceptual elements has been released (Table 8.1), based on examining several use cases from different types of assessments (Scalise, Wilson and Gochyyev, 2021[2]). The framework is expected to evolve and be revised over time as more use cases are examined.

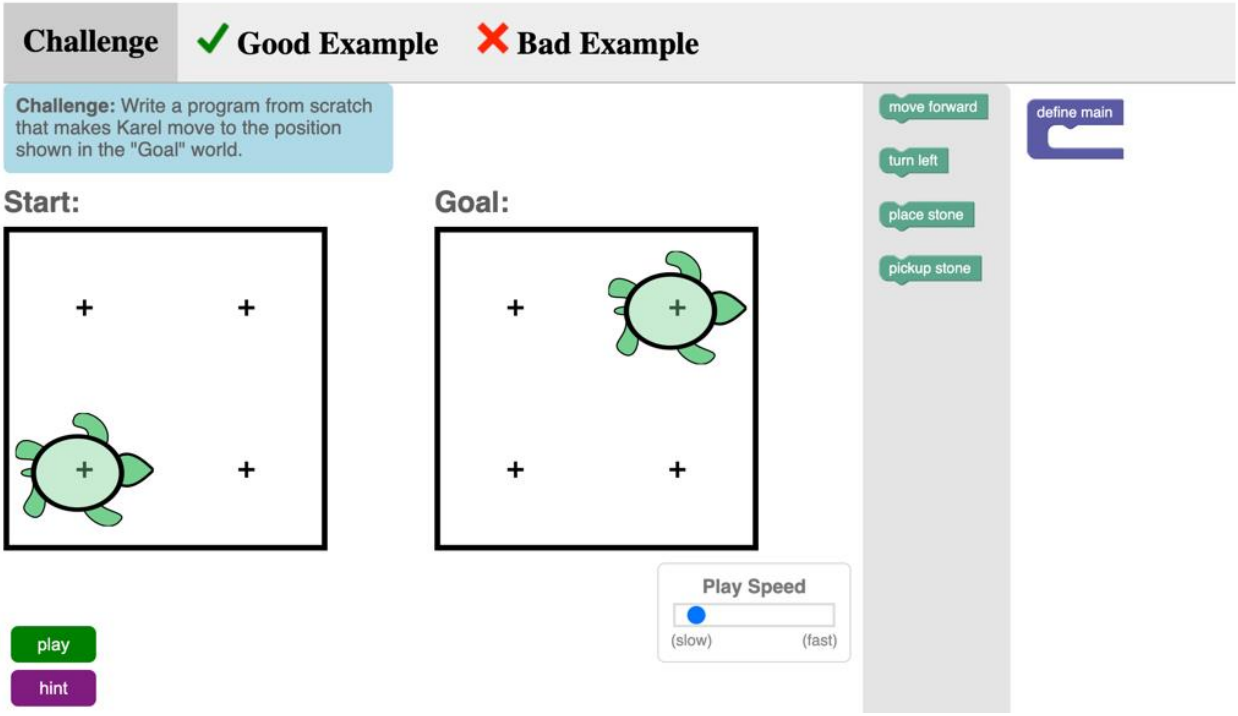
Table 8.1. A taxonomy of conceptual elements at the intersection of measurement science and learning analytics

Element	Category A (Low)	Category B (Med-Low)	Category C (Med)	Category D (High)
1: Coherence of evidentiary design	Unsupervised process data	Supervised process data	Designed process data, observation or item*	Designed process data, observation or item with framework alignments and validation
2: Self-awareness of being assessed	Stealthiness	Disclosure	Disclosure with data validation checks**	Informed consent
3: Respondent agency	Assessor selects same content for all users	Algorithms select content based on user data	Algorithms responsive to refine some content based on user choices***	User selects content
4: Focus of assessment	Analytic focus for reporting is persons	Analytic focus for reporting is groups	Analytic focus for reporting is persons/groups, with some information on items/materials	Analytic focus for reporting is items/materials
5: Open-ness	Intended goals and resulting claims reported but not methodology employed in design and analysis	Methodology reported for design but not analysis	Methodology reported for design and analysis but no technical report available	Methodology reported for design and analysis and technical report available

Notes: * Categories C and D differ as C does not align observations transparently (e.g. to a declared framework). Both use purposefully designed data rather than "found" process data not designed for the assessment purpose (e.g. social media "likes" for deducing attitudinal constructs).
** Data validation checks are when the software prompts the user to confirm their choice.
*** Overall, this refers to how much agency the user has in terms of personal choices. When the assessment developer (Category A) or computer algorithm (Category B) entirely select the content, these are lower on the spectrum of choice. Although Category B uses data based on the individual to adapt the assessment, it does not give personal choice; when the user is allowed some choice (Category C) this is considered higher on this spectrum. Full choice (Category D) is highest for this element. Recall that lower does not necessarily imply a value judgement in this taxonomy but rather describes a purpose and potentially a need for how to accumulate the evidence effectively to make an inference.
Source: Adapted from Scalise et al. (2021[2]), updated in elements 3B-C with information derived from the PILA use case.

To illustrate how this taxonomy helps to identify measurement challenges, we apply it to an assessment of computational problem solving that has recently been developed by the OECD for its Platform for Innovative Learning assessments (PILA, https://pilaproject.org/). One PILA application (Karel) involves block-based programming of a virtual robot, "Karel", to execute actions in a grid-based environment (see Figure 8.1). Karel understands a few basic instructions such as moving forward, turning left and picking up stones, and a tutorial explains to students how to program these basic movements using the block coding language. In a typical assessment activity in PILA, students complete a set of tasks that is either assembled by their teacher or by a researcher. Within the body of work selected, the student can decide how much time to dedicate to each task and (in some cases) which task they want to complete first in the sequence. This introduces measurement challenges but also allows the PILA assessments to have high agency for the learning context. This may relate to whether relevance and intrinsic motivation are perceived by the user when engaging in the tasks, especially within the naturalistic setting. Self-directed learning is also part of the framework being addressed with such tasks.

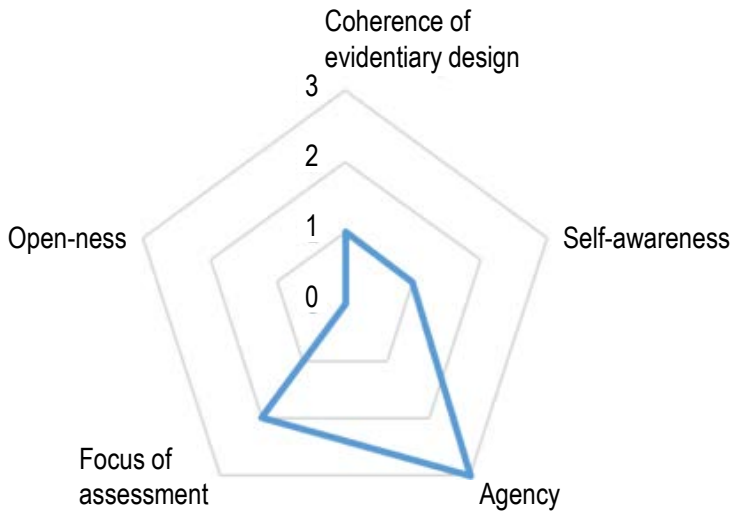
Figure 8.1. Screenshot from original version of PILA Karel task



Source: OECD (n.d.[4]), OECD's Platform for Innovative Learning Assessments (PILA), https://pilaproject.org/.

Using the five conceptual elements described in Table 8.1, the radar plot below (see Figure 8.2) illustrates where the original PILA Karel task fell in this conceptual space at the time reviewed. Note that where a task falls in the conceptual space described in Table 8.1 is not intended as a value judgement: good activities might potentially fall anywhere in the space and still offer useful evidence, but different parts of the space involve different challenges for making measurement claims. For instance, a task high in agency (such as the PILA Karel task), often considered a valuable trait by learning scientists, means the student self-selects many choices in how the task plays out. They can decide, for example, to consult a worked example or access a hint at any point in the task. This agency provides opportunities to measure self-regulated learning, which is a measurement target across all PILA activities. However, such choices may make evidence streams more or less different between respondents, and this poses measurement challenges for comparability.

Figure 8.2. Conceptual map of original version of PILA Karel task



Exploring one avenue: Borrowing strengths from another analytic technique or using one model to extend another

In the remainder of this chapter, we elaborate on the approach "borrowing strength across models to effectively aggregate evidence" as a possible solution to some of the measurement challenges discussed in the context of complex, interactive and innovative tasks. Since results can be examined for robust traits, this approach has become a common suggestion for complex technology-based experiences that want to make measurement claims. An AI group in Hong Kong, for instance, recently advocated for "Deep-IRT" (Yeung, 2019[5]):

> *"We propose Deep-IRT which is a synthesis of the item response theory (IRT) model and a knowledge tracing model that is based on the deep neural network architecture called dynamic key-value memory network (DKVMN) to make deep learning based knowledge tracing explainable. Specifically, we use the DKVMN model to process the student's learning trajectory and estimate the item difficulty level and the student ability over time. Then, we use the IRT model to estimate the probability that a student will answer an item correctly using the estimated student ability and the item difficulty. Experiments show that the Deep-IRT model retains the performance of the DKVMN model, while it provides a direct psychological interpretation of both students and items."*

Such an approach is an example of a nested or hybridised measurement model. In this case, first an ML/AI model is used to process a complex learning trajectory, then an overarching IRT model is applied for interpretation. Looked at from this perspective, many types of ML/AI or other accumulation techniques that yield ordered or even nominal ways of valuing performances over a set of complex observables could yield evidence for a hybridised measurement model. For instance, scores from many "scoring engines" used in large-scale assessment already implicitly employ such a type of hybridisation. In two phases, scores are generated by machine learning of essays, for instance, then accumulated with psychometric models to make inferences along with other data on language constructs of interest. Many examples of scoring engines exist in other fields too, such as computer science.

Box 8.1. Applying hybrid models to the Harvard VPA task "There's a New Frog in Town"

The Harvard Virtual Performance Assessment (VPA) is an immersive virtual environment with the look and feel of a videogame. In the task "There's a New Frog in Town" (New Frog), each participant participates 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).

Figure 8.3. An example screen of The New Frog VPA

Source: Scalise and Clarke-Midura (2018[6]), originally accessed at http://vpa.gse.harvard.edu/.

In New Frog, examinees 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 keyable 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) were 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[6]). 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. Note that the cost of reliably hand scoring constructed responses is often the white elephant in the room for costs of large-scale assessment. 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.

Learning tasks that also gather assessment evidence often employ Bayes Nets in scoring engines, which are highly flexible and easy to use. When content experts create naturalistic authentic technology tasks, such as simulations or serious gaming, they almost always incorporate many activities intended to frame small experiences that are easily described in a small Bayes' net. Small Bayes' nets are easy to define and to generate scores, assuming tasks have sufficient "left over" data considered salient to the construct but not yet connected together to incorporate in a formal scoring model.

One hybridised model that has appeared is mIRT-bayes (Scalise, 2017[7]). It employs small Bayesian networks to help score over rich patterns of evidence, then uses a multidimensional IRT (mIRT) model to accumulate scores and yield inferences. In mIRT-bayes, the overarching IRT model can be uni- or multidimensional and will yield the solution needed to meet the "Levy challenges" at large scale. Applied originally to simulation-based data from Harvard's Virtual Performance Assessments (VPA) (Dede, Clarke-Midura and Scalise, 2013[8]), mIRT-bayes was examined across two VPA tasks with similar results (see Box 8.1).

Although estimating the model all at once (one phase) is mathematically possible for mIRT-bayes, the two phases intentionally offer modularity. This preserves the flexibility of Bayes Nets for task design while retaining the robust statistical properties of latent variable methods for accumulation and inferences. In a modular approach, changes to the tasks are easier. If Bayes Nets are described during task development (i.e. before data collection), it can help task designers think about how to revise tasks to elicit the needed evidence. Change can also involve simply keeping and dropping scores, a well-known procedure in educational measurement, or applying simple treatments to scores without revising full analytic models

Exemplifying hybridisation with mIRT-bayes on PILA task

We exemplify a possible use of mIRT-bayes in the PILA Karel task presented in Figure 8.1. The draft scoring rubric for the task assigns credit for using control flow structures (loops and conditionals) to address repeating patterns in code. High performance for the age group involves solving a complex problem by writing nested repetition(s) of commands. Whether or not respondents can finally generate this type of code is revealed by the end product, which can be scored for "correctness". The product might be scored acceptable or not or might be given credit for partial solutions with a polytomous rubric. However, scoring only the final work product ignores considerable construct-relevant information in the process data, or interactions in which respondents engage.

For this task, process data is deemed salient by learning scientists in many screens of the PILA activity. Usually learning scientists who are experts in a given area have been selected to design a task because they can describe their ideas of what salient actions are. In other words, they often hold an implicit theory for why they designed the task as they did. Descriptions of Bayes subnets generally start with this theory. Alternatively, or to further develop implicit theories, data sets can reveal interesting patterns to improve nets. This is not unique to Bayes and is also true of other cluster and AI/ML techniques.

Because the PILA tasks allow considerable agency and incorporate a great deal of complexity through a naturalistic flow, many single actions taken by a respondent can't be expected to say much alone but may be deemed salient over a group of observables. From a measurement perspective, these groups of observations are known as "semi-amorphous" data. These data are also "semi-structured" from a computer science perspective since the observations to be captured are tagged in the data collection file and can be parsed but are not valued with a score.

With semi-amorphous data, trying to accumulate individual actions directly within an IRT model is often a validity threat. If only *sequences or patterns* over the actions are salient, each individual action alone is not salient – in other words, other equally meaningful actions could have been taken. Stakeholders often rightly reject the approach of trying to make claims on individual actions and patterns can be hard to evaluate without models to help with the process. However, if larger patterns of actions are deemed meaningful by

the learning scientists designing the tasks, then this information should be relevant to accumulate and might be included in an IRT model for reporting *if a score for the pattern is used and models can be shown to fit*. This is what hybrid models (like Deep-IRT or mIRT-bayes) are intended to accomplish. Patterns of salient actions can be identified before data collection (from theory), or after data collection if the necessary data are tagged in the software and can be parsed (i.e. semi-structured). If a pattern is identified afterward rather than from theory, then results can be considered exploratory and should be subjected to expert identification and confirmation with subsequent data sets.

Table 8.2 describes the construct-relevant information available in the PILA Karel task. For the example discussed, relevant information that might be used as traditional "item" scores in a hybridised model include success in generating a simple loop and success in extending the simple loop to a nested loop. However, some patterns to be scored on this construct are not this simple. For example, one high level of the rubric on the nested loop goal requires adapting control flow structures to generalise across multiple problems. Salient semi-amorphous information from process data might include whether available examples were accessed, whether the hint was accessed, the degree and type of prior flawed attempts to generate both simple and nested loops, and whether feedback was viewed with indicators such as NT10. The NT10 metric describes whether students took enough time to use the information and is an indicator of test effort (Wise et al., 2004[9]). In this case, it might reveal extremes of insufficient time allotted to use feedback.

Table 8.2. Construct-relevant information available in original version of PILA Karel task for accumulation on the 'nested loop' goal

Interaction or product
Traditional score possible
Success in repeated nested loop (IRT only y/n or polytomous by rubric)
Success in prior not nested repeat loop (IRT only y/n or polytomous by rubric)
Potentially other questions and items
Bayes Net node/arc possible subnet accumulations
Good example accessed (GE; y/n)
Bad example accessed (BE; y/n)
Hint accessed (hint; y/n)
Prior run with incorrect use of nested block (run; accumulated)
Total number prior runs without nested block (run; accumulated)
Feedback NT10 on any resource use listed above (NT10; y/n)
Feedback NT10 on any prior incorrect run
Run number
Unique runs
Prior knowledge subnet
Prior knowledge subnet (accumulated):
Prior knowledge survey question (y/n)
Success on first attempt for each screen (y/n)
Tutorial NT10 for attempts on each screen (y/n)

Note: y=yes, n=no.

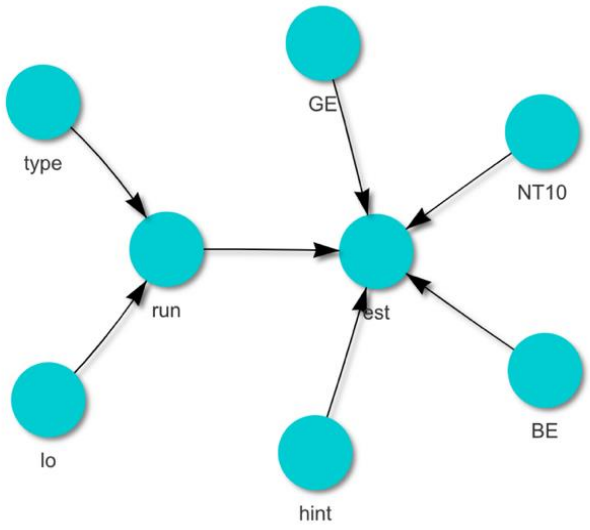
An example Bayes subnet showing a simple theoretical relationship for a few of the nodes from Table 8.2 for the PILA Karel task is shown in Figure 8.4. A subnet is a small part of a Bayes net that generates a score over a pattern as an "estimate". Often it is possible to identify several subnets to expand salient evidence used from a complex task. Theoretical ideas for how the nodes might be connected in a Bayes subnet for Figure 8.4 were based on information in the description of the task and conversations about implicit theories for how the task was designed. The elements within the subnet are labelled as follows:

- "GE" for good example accessed (yes/no);
- "BE" for bad example accessed (yes/no);

- “hint” for hint accessed (yes/no);
- “run” for prior run with incorrect use of nested block (yes/no);
- “NT10”, a flag discussed earlier for insufficient test effort based on time taken compared to normative time (0/1, if less than 10 percent of the normative time);
- “type”, a set of strategies identified *a priori* by learning scientists coded as to which ones the student tried; and
- “lo”, a flag for the total number of prior runs without the nested block (0/1).

This subnet could be applied to parsed process data on the first nested loop challenge in the PILA Karel task. The data set can be used to determine the full joint probability and query individual estimates for the subnet.

Figure 8.4. One example theoretical Bayes subnet for accumulating information from original PILA Karel task example



Source: Diagram generated through the R Bayes Net package bnlearn (Scutari, 2010[10]).

To illustrate how implicit theories can evolve into questions for data mining, there are several questions involving the “run” node of the subnet. Currently the run node is shown conditioned on the type (“type”) and number (“lo”) node parents shown in Figure 8.4. Based on the eventual data, do the node parents provide helpful information or is the dichotomised “run” node alone enough? If “lo” is informative, should we maintain intervals (e.g. 0, 1, 2, 3, 4+) or can we dichotomise to low and high? If so, where is the most informative location for the cut point? Another question relates to the “type” node: can we categorise the student exemplars of prior runs without a nested block into a set of meaningful types? In order to answer these questions, we need to examine the results in the data set and likely also consult with the task’s learning scientist developers to see if they observe patterns in the results.

Data sets are not yet available for PILA tasks but are expected to be used soon to explore if this subnet configuration is informative or if other configurations might better fit, as well as to investigate other possible subnets. This approach blends theory with data mining. Ultimately multiple data sets will be available in PILA so we also intend to compare the utility of different clustering algorithms within the mIRT model. This can help us say whether Bayes subnets are most useful or if another emerging clustering algorithm might be useful.

Such embedded model comparisons through OECD-sponsored innovations in learning assessments represent an important opportunity for innovation. For the moment, IRT itself seems a given within the context of large-scale assessment due to its ability to meet the “Levy challenges” discussed earlier. For PILA, mIRT is a suggested approach given the intended multidimensional nature of the draft framework. Recently, engines have also been developed that derive Bayes Nets automatically from task data sets and “discover” subnets that may provide additional information to a measurement model. Such subnets would still need to be reviewed for saliency by learning scientists, but this automated approach to hybridising models also poses interesting possibilities.

Conclusion

Digital technologies hold promise for helping to bring about changes in educational measurement and assessment. As described in the Introduction chapter of this report, one challenge to be faced is how to accumulate evidence in defensible ways when constructs and observations include inherent complexity. Several different approaches have been discussed and an example developed.

We argue that viable approaches should propose a set of potential solutions to the dilemmas described in this chapter. We have shown there is space to consider change and we have exemplified some approaches – but we do not intend to specify what form such change should take. We have illustrated using a PILA example for an approach to hybridise models by borrowing strength across different emerging and traditional methods. Hybrid models may flexibly accumulate patterns of evidence and generate scores using ML/AI approaches such as Bayes Nets or more current clustering algorithms, while preserving an overarching IRT model to meet the “Levy challenges”. Such innovations might help not only to empower the use of technology affordances discussed in earlier chapters of this report, but also could generate important advances in measurement technologies for hard-to-measure constructs and novel, engaging activities for students.

References

Dede, C., J. Clarke-Midura and K. Scalise (2013), "Virtual performance assessment and games: Potential as learning and assessment tools", in *Paper presented at the Invitational Research Symposium on Science Assessment, Washington, D.C.*. [8]

Levy, R. (2012), "Psychometric advances, opportunities, and challenges for simulation-based assessment", *Invitational Research Symposium on Technology Enhanced Assessments*, ETS, Washington, D.C., https://www.ets.org/Media/Research/pdf/session2-levy-paper-tea2012.pdf. [3]

OECD (n.d.), *Platform for Innovative Learning Assessments*, https://pilaproject.org/ (accessed on 3 April 2023). [4]

Papamitsiou, Z. and A. Economides (2014), "Learning analytics and educational data mining in practice: A systemic literature review of empirical evidence", *Educational Technology and Society*, Vol. 17/4, pp. 49-64. [1]

Scalise, K. (2017), "Hybrid measurement models for technology-enhanced assessments through mIRT-bayes", *International Journal of Statistics and Probability*, Vol. 6/3, pp. 168-182, https://doi.org/10.5539/ijsp.v6n3p168. [7]

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. [6]

Scalise, K., M. Wilson and P. Gochyyev (2021), "A taxonomy of critical dimensions at the intersection of learning analytics and educational measurement", *Frontiers in Education*, Vol. 6, https://doi.org/10.3389/feduc.2021.656525. [2]

Scutari, M. (2010), "Learning Bayesian Networks with the bnlearn R package", *Journal of Statistical Software*, Vol. 35/3, https://doi.org/10.18637/jss.v035.i03. [10]

Wise, S. et al. (2004), "An investigation of motivation filtering in a statewide achievement testing program", in *Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, California*. [9]

Yeung, C. (2019), "Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory", in *EDM 2019 - Proceedings of the 12th International Conference on Educational Data Mining, Montréal, Canada*, https://doi.org/10.48550/arXiv.1904.11738. [5]

9 Measuring self-regulated learning using feedback and resources

By Ido Roll and Miri Barhak-Rabinowitz

(Technion – Israel Institute of Technology)

This chapter describes an assessment of self-regulated learning (SRL) that is based on an extended learning task with embedded feedback and resources. These resources serve the dual purpose of providing learners with meaningful choices to support their learning as well as documenting their behaviours. Three affordances of such resources are identified: experimentation, where learners can express and test ideas; explicit feedback, where learners can monitor their progress towards their learning goals; and information-seeking, where learners can receive information about the task and its environment. The chapter analyses the PISA 2025 Learning in the Digital World assessment to demonstrate how these resources support the assessment of SRL. The chapter then addresses the interplay between data and theory in constructing learning and assessment tasks. It discusses some of the main design challenges, focusing particularly on the effect of prior knowledge and supporting generalisable inferences.

Introduction

The goals of assessment are expanding from evaluating the application of existing static knowledge (learning *outcomes*) to evaluating the dynamics of acquiring and developing new knowledge (learning *processes*) – see Chapters 2 and 4 of this report as well as Bransford and Schwartz (1999[1]), National Research Council (2001[2]) and Cutumisu, Chin and Schwartz (2019[3]). To support such inferences, assessment should provide learners with authentic and complex problem-solving tasks where learners are required to manage their own learning and exercise agency. Indeed, common to most 21st Century skill frameworks is the view of students as agentic learners who regulate their learning process (Dede, 2010[4]; Kirschner and Stoyanov, 2018[5]; OECD, 2018[6]). This chapter takes a deeper dive into how assessment can support inferences about agentic learners and discusses the affordances, design guidelines as well as challenges of using digital resources to assess students' self-regulated learning (SRL).

The term SRL refers to students' goal-directed actions that guide the dynamics of accessing, constructing, and applying knowledge (Schunk and Zimmerman, 2013[7]). SRL includes the use of cognitive and metacognitive strategies as well as regulating one's motivational and affective states (Panadero, 2017[8]). Cognitive strategies support making progress towards one's learning goals. Metacognitive strategies support the planning, monitoring and adjustment of these strategies by setting goals, monitoring progress and adjusting the learning process (Flavell, 1979[9]; Schunk and Zimmerman, 2013[7]). The motivational and affective components of SRL refer to the processes through which learners manage their emotional states while learning such as their willingness to persist with learning activities and environments even in the face of difficulty (Fredricks, Blumenfeld and Paris, 2004[10]; Järvenoja et al., 2018[11]; OECD, forthcoming[12]).

As implied by this view of SRL, regulative processes are essential in authentic learning contexts that have the following characteristics: 1) a challenge to overcome such as a goal to be learned or a problem to be solved; 2) the challenge is non-linear requiring learners to make meaningful choices that affect how they progress towards their goal or solution. Providing learning resources serves two important goals in this context. First, they support non-linear learning trajectories and invite learners to make meaningful choices. Second, they collect digital traces of these choices. Learning resources in complex tasks thus offer valuable opportunities to assess SRL (Roll and Winne, 2015[13]; Shute and Rahimi, 2021[14]).

We begin this chapter by describing key affordances of digital learning resources that facilitate and capture SRL behaviours. We then demonstrate the use of such resources in assessment using the example of the PISA 2025 Learning in the Digital World assessment (OECD, forthcoming[12]). Last, we describe design considerations and main challenges for inferring SRL skills from digital trace data.

Resources and affordances that support the assessment of SRL

Providing learners with digital tools and representations is a known strength of learning with technology. Digital tools that help learners organise their thinking and support knowledge construction are often referred to as "cognitive tools" (Drew, 2019[15]; Jonassen, 1992[16]; Nesbit, Niu and Liu, 2018[17]). We view digital learning resources as a class of cognitive tools *that offer interactivity* to support meaning-making. We emphasise interactivity not only in that learners can use these tools flexibly but also in that the tools provide learners with additional information that is not available without the tools. Being interactive, resources provide learners with opportunities to access new knowledge. Being digital, each action that learners make can be logged, thus offering a window into their reasoning and learning processes.

Learning resources offer learners multiple affordances to enact goal-oriented behaviours. We group these 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 outcomes (conversely, coding tasks where learners enter code but cannot execute it are not considered learning resources here). Another example is interactive scientific simulations where learners can manipulate elements and observe the outcome of their exploration (Wieman, Adams and Perkins, 2008[18]). The main benefit of such resources comes from their responses to learner actions, often termed "situational feedback" (Nathan, 1998[19]; Roll et al., 2014[20]). Situational feedback provides learners with representations of the real-world equivalence that follows from learners' actions. For example, an interactive simulation for electricity will adjust the shown light intensity based on the voltage that learners set (Roll et al., 2018[21]; de Jong et al., 2018[22]). Situational feedback is implicit and originates within the task situation itself, consistent with the internal logic of the task. That is, learners are not flagged or graded by an external all-knowing model but instead given opportunities to elicit, observe and interpret relevant information from the environment response (Nathan, 1998[19]). 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 enable learners to evaluate 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[23]). 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 like monitoring or which sub-goals they pursue (Winstone et al., 2016[24]). As with situational feedback, students who adjust their cognitive strategies effectively following explicit feedback demonstrate productive metacognition (Kinnebrew, Segedy and Biswas, 2017[25]).

Information-seeking affordances support learners by providing additional communication about the task at hand. Informational resources include hints (Aleven et al., 2016[26]), instructional videos (Seo et al., 2021[27]), worked examples (Ganaiem and Roll, 2022[28]; Glogger-Frey et al., 2015[29]), searchable databases, etc. Information sources can be fixed, as in most tutorials, or adaptive, as in hints about the specific problem step (see VanLehn et al. (2007[30]) for example). 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[27]; Wood, 2001[31]). Thus, interactions with information resources can provide meaningful insights into learners' help-seeking and monitoring processes (Roll et al., 2014[20]).

Table 9.1 summarises these three families of learning affordances and their associated aspects of SRL.

Table 9.1. Affordances of digital learning resources and opportunities to elicit and evaluate SRL

Affordance	Description	Opportunities for SRL assessment	Examples of resources
Experimentation	Enabling students to express and evaluate different ideas.	Evaluating the enactment of sub-goals and use of affordances to adjust strategy use.	• Coding environments • Executable concept maps • Interactive simulations
Explicit feedback	Allowing learners to evaluate their progress towards their goal.	Evaluating learners' use of feedback to reduce uncertainty and monitor progress.	• "Test" button to check solution correctness • Automatic feedback in the form of error flagging or explanations
Information-seeking	Allowing learners to access and curate new information on demand.	Evaluating learners' effectiveness in identifying knowledge gaps and choice to seek information to support problem solving.	• Tutorials • Searchable information sources • On-demand hints • Worked examples/contrasting cases

Tracking evidence for SRL using resources: The PISA 2025 Learning in the Digital World assessment

As described in the Introduction and Chapter 5 of this report, Evidence Centred Design (ECD) provides a principled framework for designing digital assessments of complex constructs. It can therefore support the design of task features and affordances that elicit relevant evidence about SRL-related target competencies. At the heart of the ECD framework for SRL are rules that associate observed test behaviours (evidence) with the target SRL competencies (inferences). ECD breaks this process into two types of rules: 1) evidence rules, which quantify observations about learners' outputs (or SRL behaviours, in our case); and 2) a statistical model, which specifies the relationship between these observations and estimates of learner competencies (Mislevy, 2013[32]). Here we use the term "rules" to describe the combination of these processes into a single evidence model that links observable behaviours with inferred SRL competencies. We focus on the PISA 2025 Learning in the Digital World (LDW) assessment to demonstrate the use of such rules to design tasks for assessment of SRL.

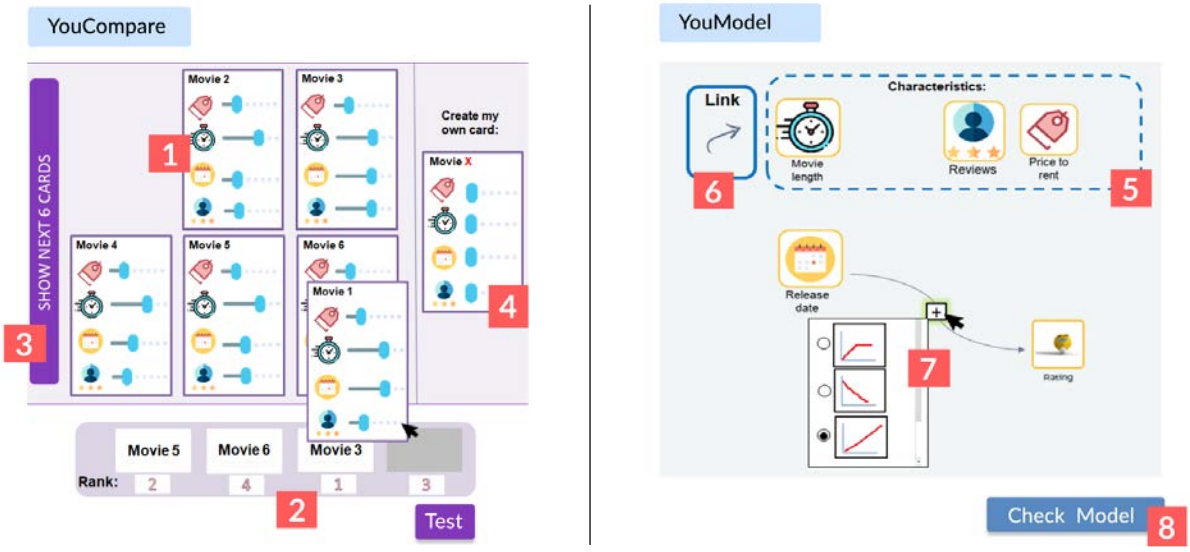
Chapter 6 of this report elaborated on the assessment design processes of domain analysis and domain modelling in the context of the LDW assessment, as first steps in a principled ECD process. As a reminder, PISA 2025 defines the LDW construct as "the capacity to engage in an iterative process of knowledge building and problem solving using computational tools. This capacity is demonstrated by effective self-regulated learning while applying computational and scientific inquiry practices" (OECD, forthcoming[12]). Here, we focus on the SRL-related components of the task and evidence models for the assessment to exemplify how feedback and resources can provide inferences about test takers' SRL competencies.

In one prototype unit for this assessment named "I Like That!", described in the draft framework document (OECD, forthcoming[12]), learners are asked to create a recommender system that evaluates movie properties and predicts their popularity with a certain user called Alex. While the unit is concerned with movie preferences, it is, in fact, a scientific inquiry task. Learners investigate the relationship between predictive variables (such as price, length, release data and reviews) and the outcome variable (Alex's movie preferences).

The "I Like That!" prototype unit includes several resources that support the evaluation of learners' SRL. One key resource is an interactive data inquiry tool. Using this tool, aptly named "YouCompare", learners can compare features of different movies to identify their underlying relationship with Alex's viewing preferences (see Figure 9.1). Each movie is represented using a card (see [1] in Figure 9.1). Learners can then choose different cards and compare their attributes and their rankings on the testbed (see [2]). Students can also ask to see additional cards (see [3]). This process is analogous to choosing specific experimental set-ups to compare. Thus, YouCompare affords *experimentation*. The movies were designed so that learners can study each property (such as price or length) in isolation as well as interactions between properties. Productive learners are expected to use various cognitive strategies such as the control of variables strategy (CVS) to evaluate these effects. For example, learners can compare cards with different movie lengths and the same characteristics for all the other variables to identify the relationship between movie length and Alex's preferences. Furthermore, learners can also create test cases to compare with the provided cards, specifying the values of the variables for that card by adjusting the sliders (see [4]).

A second interactive resource of the task is "YouModel". This resource allows learners to create models for their recommender system. Learners first identify the relevant properties (see [5]) and links them to their model (see [6]). Learners then specify the relationship between the variables by choosing graphs that correspond to the desired relationship on a concept map (see [7]). When pressing the "Check Model" button, the system provides feedback on the model by highlighting elements that are incorrect (see [8]). Thus, the YouModel tool affords *explicit feedback*.

Figure 9.1. Resources in the "I Like It!" task of the PISA 2025 LDW assessment



Source: OECD (forthcoming[12]).

The LDW assessment also includes *information-seeking* resources. One way in which information is made available is via worked examples. Learners are given the option to study challenges together with their solutions. Each challenge-solution set focuses on a different aspect of the "I Like That!" task. These worked examples allow learners to study solution strategies that are applicable to the main task. From an assessment perspective, these worked examples can reveal how learners identify their own knowledge gaps and act strategically to overcome them.

Deriving evidence of SRL using resources in "I Like That!"

To make inferences regarding learners' SRL, the resources were designed in tandem with evidence rules that support the interpretation of various behavioural patterns. Table 9.2 demonstrates how different behaviours produce evidence that supports inferences about students' SRL competencies, as defined by the LDW framework document (OECD, forthcoming[12]). While this list showcases a variety of inferences that can be made from students' resource usage, for various reasons, not all rules are implemented in the specific "I Like That!" task.

The rules in Table 9.2 apply different types of indicators that serve as evidence for different aspects of SRL. Indicators can be derived from the choice to engage with a certain resource (e.g. viewing a worked example as strategic help-seeking). Indicators can also be derived from the way the resource is used (e.g. how learners navigate worked examples can provide evidence for their awareness of knowledge gaps). Finally, indicators can be derived from the actions that follow the use of the resource (e.g. applying the provided advice or information). One approach for interpreting learners' actions in context is coherence analysis, which looks for logical sequences of actions taken by the student (Kinnebrew, Segedy and Biswas, 2017[25]). Coherence analysis assumes that learners' actions generate information that can then be used in subsequent actions; when learners act on this information, then their actions are coherent. For example, the action of pause, by itself, cannot be evidence of reflection. It is important to use patterns of actions, rather than single actions, as evidence (see also Chapter 8 of this report). A coherent sequence of pauses followed by modelling of the tested relationship is a positive sign of reflection that can be interpreted as evidence. It is important to emphasise that interpretation of such patterns should be validated using data, as explained below.

Table 9.2. SRL inferences and supporting evidence

SRL inferences	Evidence
Evaluating and identifying one's knowledge gaps	• The learner asks to see the relevant worked example when stuck. • The learner consults with the correct tutorials when using a new tool.
Planning appropriate sub-goals	• The learner tests and then models one variable after the other.
Monitoring progress towards one's goals and adapting strategies	• Once the leaner asks for additional information, they apply actions that are consistent with this information (e.g. enacting the relevant strategy from a worked example). • The learner conducts frequent tests as they build and edit their computational artefacts.
Managing one's motivation and affect, so to persist in the face of difficulty	• The learner continues working after failures. • The learner uses all the available time.
Reflecting on one's performance	• The learner tests their final answer before submitting it. • The learner opts to view correct solutions after finishing sub-tasks.

Designing tasks and resources for SRL assessments

There are several considerations to keep in mind in the design of an assessment of SRL.

Designing tasks that support agency

An essential element of assessments that target the measurement of SRL is to provide student agency through choice (Bransford and Schwartz, 1999[1]; Cutumisu et al., 2015[33]). To provide meaningful agency, learners should be given a large interaction space where their choices have visible and meaningful consequences that affect the task situation. We contrast that with more constrained assessments with a pre-determined sequence of desired actions. Indeed, in the "I Like That!" prototype unit described earlier there are many ways for students to engage with the tasks, for example by isolating different variables, testing different movies, plotting relationships, etc.

The large design space for learners to explore does not mean that the goal state is under-defined. In fact, to support assessment, the goal state should be clearly defined and distance from the goal state should be quantifiable. For example, in the "I Like That!" prototype unit the correct solution is the underlying model that determines the movie ratings and overall performance can be measured by looking at deviations from this model.

Designing evidence rules

Table 9.2 describes rules for interpreting evidence of productive SRL behaviours. The converse is also true – rules can interpret evidence of ineffective regulation. Assessments of SRL can have a diagnostic value and capture common non-productive behaviours. For example, competent learners who ask for help might signal an intention to game the system.

Designing evidence rules also depends to a large degree on our model of the domain, commonly referred to as the competency model. To be able to define SRL rules, this model must be very specific (Roll et al., 2007[34]). Saying that students who struggle need help is probably true, but it is insufficient as an inference rule. Rules that include specific behaviours – such as learners who do not progress within five minutes should look at examples, as done in the "I Like That!" unit – are much more informative.

A good approach for identifying relevant evidence rules (and their level of specificity) is to combine a top-down (i.e. theory first) process with a bottom-up (i.e. data first) process of knowledge discovery. Ritter and colleagues (2019[35]) describe several methods for identifying sequences of actions that yield productive learning. In the context of the "I Like It!" unit, such sequence mining can help us identify productive or efficient approaches to solutions as well as ones that were not anticipated, such as which type of information resource is useful and when. However, one should avoid the tyranny of data in which correlational evidence suffices. Each identified rule should have a strong theoretical justification that provides a mechanistic explanation for how a specific SRL construct manifests itself in the observed behaviour.

Key challenges

Generalisability

A major challenge for any complex assessment – and assessments of SRL are no different – is that of generalisability and validity (see Chapter 1 of this report for more on these challenges in assessment, more generally). While some rules may be too broad (and thus provide inconclusive evidence), others may focus on overly specific solution approaches that do not apply to other tasks or environments. For example, is clicking a button marked "?" in a specific test environment indicative of the quality of one's help-seeking in other contexts?

The challenge of generalisability goes back to how the competency model is defined (see Chapter 6 of this report for more on how the domain and competency model for the LDW assessment were defined). The competency model aims to define aspects of SRL, often considered a rather domain-general set of competencies. However, SRL behaviours are only meaningful in context and can only be interpreted within the specific context in which they occur (as argued in Chapter 2 of this report). Simon (1969[36]) describes the journey of an ant on a sandy beach and how interpreting the ant's behaviour should consider the terrain of the beach. This creates a major challenge for learning about ant behaviour, and similarly, for learning about students' SRL competencies. What are reasonable boundaries of generalisation from the instantiated behaviours in the assessment? How can tasks be designed to support such generalisation?

Several solution approaches can mitigate this challenge. The first is to triangulate inferences using several rules to infer about each construct. As seen in Table 9.2, SRL constructs can use evidence from multiple rules. Notably, when these rules are applied in the same task, they do not solve the dependency of the observations on the task topic and scenario. Another solution is to design tasks that use parallel evidence rules. Constructs that are assessed in "I Like That!" are also assessed, for each student, in at least one additional task with a different topic and set of tools. For instance, instead of a concept map for experimentation, alternative tasks use a block-based coding tool. This is similar to observing ants on multiple beaches. This approach also helps to distil key features of the task model (that are similar across tasks) from more superficial ones (that can vary across tasks).

An important risk when designing evidence rules is construct-irrelevant variance. For instance, productive use of verbose hints may be indicative of reading comprehension more than of help-seeking and vice versa (unproductive use of these resources may indicate a lack of reading comprehension). In fact, construct-irrelevant variance may lead to opposite results than intended. For example, when hints are too verbose or unhelpful, students who regulate their learning well typically avoid help (Roll et al., 2014[20]). A related challenge to validity is cultural relevance, as rules may unintentionally introduce biases due to different ways of approaching challenges across cultural groups (see also Chapter 11 of this report).

As mentioned above, one approach to examining the generalisability and validity of rules relies on the interplay between theory and data. It is essential to collect data early on to validate the rules. The provision

of think-aloud protocols and cognitive labs is effective in that regard. A good evidence model has both predictive power (who learns well?) and explanatory power (what makes them learn well?).

The role of prior knowledge

Numerous attributes mediate the relationship between SRL and the use of feedback and resources. Identifying these attributes and assessing performance as a function of those variables is a significant challenge for interpreting student behaviours. One key dependency is that the strategic use of resources depends on students' domain knowledge, as experts and novices manage their learning differently (Kalyuga and Singh, 2015[37]; Zohar and Barzilai, 2013[38]).

SRL assessments should consider two aspects of prior knowledge. One aspect is domain-level knowledge – for instance, knowing and understanding a domain inevitably affects one's need for help. The second aspect is familiarity with relevant resources – for example, students who are used to concept maps or learning from examples may have an advantage when encountering these resources. The coupling between SRL behaviours and prior knowledge of the topic or tools is inherent to any task, as students regulate their learning to overcome knowledge gaps. Thus, knowledge gaps are a key motivation for choosing which SRL strategies to enact.

One approach to mitigate the effect of prior knowledge is to assess it and adapt rules accordingly. This means that rules become conditional on prior knowledge – for example, asking for help *only* when lacking knowledge or trying to correct errors on their own *if* sufficiently knowledgeable. In some cases, assessing prior knowledge is rather straightforward: when tasks build on well-defined domains such as coding or mathematics, traditional items can be used to assess domain knowledge. However, as most SRL assessments require complex tasks, assessing prior knowledge is non-trivial. For example, in the "I Like That!" prototype unit, assessing prior knowledge of modelling practices is challenging. One solution is deconstructing and testing each modelling practice separately, without resources. A similar approach can be used to evaluate prior knowledge of tools. For example, learners can be asked to perform specific tasks with the tools to assess technical fluency, although such an approach is inefficient and unauthentic.

Another approach is to minimise the relevance of learners' prior knowledge in the task. One good example of this approach is the "I Like That!" prototype. This task is a typical scientific inquiry scenario: students are given data and are asked to identify the relationship between different variables by applying inquiry strategies (Pedaste et al., 2015[39]; Roll et al., 2018[21]). However, implementing the task in a scientific context would have created construct-irrelevant variance, namely prior exposure to the relevant scientific topics. Instead, the task uses a made-up scenario. This scenario levels the playing field in numerous ways. Intuitively, most students understand what a recommender system for movies (such as Netflix) does. Practically, though, very few have experimented with building one. In addition, learners do not have prior knowledge of the model of the specific made-up user.

A third approach seeks to minimise variability in prior knowledge. By using tutorials, examples and walkthrough problems, students can be given basic knowledge with which they can approach the task. This is especially useful regarding knowledge of the task-specific tools, such as learning to use the concept map tool in the "I Like That!" example.

Motivation regulation

This chapter focused on assessing metacognitive and cognitive strategies using feedback and resources. Such analysis overlooks the very significant role of motivational regulation in SRL. Prior work has demonstrated the potential of assessing affective states in digital environments (Calvo and D'Mello, 2010[40]; Woolf et al., 2009[41]). The provision of resources may further aid this assessment. For example, students' use of hints can be used to identify learners who attempt to make progress without putting in the required effort (Baker et al., 2013[42]). However, the interaction between affective states and metacognitive

strategies should be further studied, especially regarding resource use (Shum and Crick, 2012[43]). Using resources requires deliberation and sustained effort (Tishman, Jay and Perkins, 1993[44]). Students need to be aware of their motivational states, their ability to control them and their impact on their learning processes (Wolters, 2003[45]). This in turn affects learners' engagement (or disengagement) (Miele and Scholer, 2017[46]; O'Brien et al., 2022[47]). Similarly, additional data sources such as self-reports may be warranted. However, such analysis is beyond the scope of the current chapter.

Conclusion

The assessment of SRL requires open and interactive task situations in which learners have agency through choice. The availability of interactive learning resources provides ample opportunities to assess the way learners make these choices. Capitalising on students' choices to assess SRL is done using an inference model, triangulated across tasks, validated using data and contingent on prior knowledge. Such assessments are exciting in that they measure students' capacity to learn, above and beyond their static knowledge state.

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. [26]

Baker, R. et al. (2013), "Modeling and studying gaming the system with educational data mining", in *International Handbook of Metacognition and Learning Technologies*, Springer International Handbooks of Education, Springer, New York, https://doi.org/10.1007/978-1-4419-5546-3_7. [42]

Bransford, J. and D. Schwartz (1999), "Rethinking transfer: A simple proposal with multiple implications", *Review of Research in Education*, Vol. 24/1, pp. 61-100, https://doi.org/10.3102/0091732X024001061. [1]

Calvo, R. and S. D'Mello (2010), "Affect detection: An interdisciplinary review of models, methods, and their applications", *IEEE Transactions on Affective Computing*, Vol. 1/1, pp. 18-37, https://doi.org/10.1109/t-affc.2010.1. [40]

Cutumisu, M. et al. (2015), "Posterlet: A game-based assessment of children's choices to seek feedback and to revise", *Journal of Learning Analytics*, Vol. 2/1, https://doi.org/10.18608/jla.2015.21.4. [33]

Cutumisu, M., D. Chin and D. Schwartz (2019), "A digital game-based assessment of middle-school and college students' choices to seek critical feedback and to revise", *British Journal of Educational Technology*, Vol. 50/6, pp. 2977-3003, https://doi.org/10.1111/bjet.12796. [3]

de Jong, T. et al. (2018), "Simulations, games, and modeling tools for learning", in Fischer, F. et al. (eds.), *International Handbook of the Learning Sciences*, Routledge, New York, https://doi.org/10.4324/9781315617572-25. [22]

Dede, C. (2010), "Comparing frameworks for 21st century skills", in Bellanca, J. and R. Brandt (eds.), *21st Century Skills*, Solution Tree Press, Bloomington. [4]

Deeva, G. et al. (2021), "A review of automated feedback systems for learners: Classification framework, challenges and opportunities", *Computers & Education*, Vol. 162, pp. 1-43, https://doi.org/10.1016/j.compedu.2020.104094. [23]

Drew, C. (2019), "Re-examining cognitive tools: New developments, new perspectives, and new opportunities for educational technology research", *Australasian Journal of Educational Technology*, Vol. 35/2, https://doi.org/10.14742/ajet.5389. [15]

Flavell, J. (1979), "Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry", *American Psychologist*, Vol. 34/10, p. 906. [9]

Fredricks, J., P. Blumenfeld and A. Paris (2004), "School engagement: Potential of the concept, state of the evidence", *Review of Educational Research*, Vol. 74/1, pp. 59-109, https://doi.org/10.3102/00346543074001059. [10]

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*, pp. 727-732. [28]

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. [29]

Järvenoja, H. et al. (2018), "Capturing motivation and emotion regulation during a learning process", *Frontline Learning Research*, Vol. 6/3, pp. 85-104, https://doi.org/10.14786/flr.v6i3.369. [11]

Jonassen, D. (1992), "What are Cognitive Tools?", in *Cognitive Tools for Learning*, Springer, Berlin/Heidelberg, https://doi.org/10.1007/978-3-642-77222-1_1. [16]

Kalyuga, S. and A. Singh (2015), "Rethinking the boundaries of cognitive load theory in complex learning", *Educational Psychology Review*, Vol. 28/4, pp. 831-852, https://doi.org/10.1007/s10648-015-9352-0. [37]

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. [25]

Kirschner, P. and S. Stoyanov (2018), "Educating youth for nonexistent/not yet existing professions", *Educational Policy*, Vol. 34/3, pp. 477-517, https://doi.org/10.1177/0895904818802086. [5]

Miele, D. and A. Scholer (2017), "The role of metamotivational monitoring in motivation regulation", *Educational Psychologist*, Vol. 53/1, pp. 1-21, https://doi.org/10.1080/00461520.2017.1371601. [46]

Mislevy, R. (2013), "Evidence-centered design for simulation-based assessment", *Military Medicine*, Vol. 178/10S, pp. 107-114, https://doi.org/10.7205/milmed-d-13-00213. [32]

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. [19]

National Research Council (2001), *Knowing What Students Know*, National Academies Press, Washington, D.C., https://doi.org/10.17226/10019. [2]

Nesbit, J., H. Niu and Q. Liu (2018), "Cognitive tools for scaffolding argumentation", in Adesope, O. and A. Rud (eds.), *Contemporary Technologies in Education: Maximising Student Engagement, Motivation and Learning*, Springer, Cham, https://doi.org/10.1007/978-3-319-89680-9_6. [17]

O'Brien, H. et al. (2022), "Rethinking (dis)engagement in human-computer interaction", *Computers in Human Behavior*, Vol. 128, pp. 107-109, https://doi.org/10.1016/j.chb.2021.107109. [47]

OECD (2018), *The Future We Want*, https://www.oecd.org/education/2030-project/contact/E2030%20Position%20Paper%20(05.04.2018).pdf. [6]

OECD (forthcoming), *PISA 2025 Learning in the Digital World assessment framework (draft)*, OECD Publishing, Paris. [12]

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

Pedaste, M. et al. (2015), "Phases of inquiry-based learning: Definitions and the inquiry cycle", *Educational Research Review*, Vol. 14, pp. 47-61, https://doi.org/10.1016/j.edurev.2015.02.003. [39]

Ritter, S. et al. (2019), "Identifying strategies in student problem solving", in Sinatra, A. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems*, US Army Research Laboratory, Orlando. [35]

Roll, I. et al. (2007), "Designing for metacognition—applying cognitive tutor principles to the tutoring of help seeking", *Metacognition and Learning*, Vol. 2/2-3, pp. 125-140, https://doi.org/10.1007/s11409-007-9010-0. [34]

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. [21]

Roll, I. et al. (2014), "Tutoring self- and co-regulation with Intelligent Tutoring Systems to help students acquire better learning skills", in Sottilare, R. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems*, US Army Research Laboratory, Orlando. [20]

Roll, I. and P. Winne (2015), "Understanding, evaluating, and supporting self-regulated learning using learning analytics", *Journal of Learning Analytics*, Vol. 2/1, pp. 7-12, https://doi.org/10.18608/jla.2015.21.2. [13]

Schunk, D. and B. Zimmerman (2013), "Self-regulation and learning", in Reynolds, W. and G. Miller (eds.), *Handbook of Psychology*, John Wiley & Sons, Hoboken. [7]

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. [27]

Shum, S. and R. Crick (2012), "Learning dispositions and transferable competencies: Pedagogy, modelling and learning analytics", *2nd International Conference on Learning Analytics & Knowledge*, http://oro.open.ac.uk/32823/1/SBS-RDC-LAK12-ORO.pdf. [43]

Shute, V. and S. Rahimi (2021), "Stealth assessment of creativity in a physics video game", *Computers in Human Behavior*, Vol. 116, pp. 1-13, https://doi.org/10.1016/j.chb.2020.106647. [14]

Simon, H. (1969), *The Sciences of the Artificial*, The MIT Press. [36]

Tishman, S., E. Jay and D. Perkins (1993), "Teaching thinking dispositions: From transmission to enculturation", *Theory Into Practice*, Vol. 32/3, pp. 147-153, https://doi.org/10.1080/00405849309543590. [44]

VanLehn, K. et al. (2007), "What's in a step? Toward general, abstract representations of tutoring system log data", in Conati, C., K. McCoy and G. Paliouras (eds.), *User Modelling 2007. Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg. [30]

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. [18]

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. [24]

Wolters, C. (2003), “Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning”, *Educational Psychologist*, Vol. 38/4, pp. 189-205, https://doi.org/10.1207/s15326985ep3804_1. [45]

Wood, D. (2001), “Scaffolding, contingent tutoring and computer-supported learning”, *International Journal of Artificial Intelligence in Education*, Vol. 12, pp. 280-292. [31]

Woolf, B. et al. (2009), “Recognizing and responding to student affect”, in Jacko, J. (ed.), *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction, Lecture Notes in Computer Science*, Springer, Berlin/Heidelberg, https://doi.org/10.1007/978-3-642-02580-8_78. [41]

Zohar, A. and S. Barzilai (2013), “A review of research on metacognition in science education: Current and future directions”, *Studies in Science Education*, Vol. 49/2, pp. 121-169, https://doi.org/10.1080/03057267.2013.847261. [38]

10 Artificial Intelligence-enabled adaptive assessments with Intelligent Tutors

By Xiangen Hu, Keith Shubeck and John Sabatini

(University of Memphis)

This chapter presents an adaptive assessment framework inspired by Intelligent Tutoring Systems (ITS), i.e. computer-based learning environments providing personalised instruction and feedback to learners through digital tutors and/or “peer” avatars. Unlike typical assessment environments, ITS provide students with interactive scenarios that integrate dynamic tasks and iterative feedback, where learning takes place as learners progress through the activities. The chapter argues that next-generation assessments have a lot to learn from ongoing developments in ITS, especially from original applications of Artificial Intelligence (AI) that provide intelligent feedback to students, adapt the system content in response to their actions, and evaluate what they know and can do. Three examples of ITS that integrate AI- applications for such purposes are presented.

Introduction

The preceding chapters in this publication have argued that assessing what students know and can do requires a coherent chain linking: 1) a vision of what relevant competencies in a given domain are; with 2) the type of tasks where observing proficiency in such competencies is possible; and 3) the appropriate methods to interpret and summarise the data resulting from these tasks. Such evidence-based reasoning is essential to make valid inferences on complex competencies, like communication and creative thinking, which people need for life and work in the 21st Century. It has been argued that eliciting evidence of 21st Century competencies requires immersive, realistic assessment tasks where test takers engage these competencies and where evaluators pay as much attention to the process by which learners come to solve the given task as to the result of this process. Finally, it has also been argued that making sense of the complex data streaming from such interactive tasks is only possible at scale by leveraging modern digital technologies.

In this chapter, we elaborate on these points by introducing an adaptive assessment framework inspired by Intelligent Tutoring Systems, i.e. computer-based learning environments providing personalised instruction and feedback to learners through digital tutors and/or peers called avatars (Nwana, 1990[1]; Rus et al., 2013[2]). Intelligent Tutoring Systems (ITS) have been one of the most active areas of research and development in learning sciences and technologies in the past 35 years. Researchers in this area have always been at the forefront of technologies, particularly the assessment technologies (Sottilare et al., 2017[3]). Most ITS applications are a unique combination of learning theories, applications in the domain and available technologies.

Unlike typical assessment environments, ITS provide interactive scenarios filled with dynamic tasks and constant feedback, where learning takes place as students progress through the activities. We contend that next-generation assessments have a lot to learn from ongoing developments in ITS, and in particular, from original applications of Artificial Intelligence (AI) that provide intelligent feedback, adapt content in response to the actions of test takers, and evaluate what they know and can do. We provide a few examples of ITS for assessment along these lines before turning to a summary of the chapter's key points and concluding thoughts.

Why are ITS relevant to next-generation assessments?

In traditional tests, the goal is to assess students' acquired knowledge prior to the task. Usually no feedback is given, tasks are likely very distinct from one another, and the type of responses is mostly limited to categorical responses (i.e. correct or incorrect answers) to minimise the "testing effect", i.e. learning from the test (Avvisati and Borgonovi, 2020[4]; Butler, 2010[5]; Dempster, 1996[6]; Roediger and Karpicke, 2006[7]; Rowland, 2014[8]; Wheeler and Roediger, 1992[9]). In contrast, in ITS – as in learning environments more broadly – the goal is to maximise the testing effect to support student learning. To this end, tasks are not only related but carefully constructed feedback is provided to students after each response. Additionally, attempting to emulate what human educators do, ITS adapt and make their judgements based on a wide range of student behaviours beyond the provision of categorical responses including response latency, signs of confidence, emotions, body movement, etc.

These differences (presence or absence of feedback, relations between tasks and observed student behaviours) between the two environments result in fundamentally different assessment approaches. The fact that tasks in typical assessment environments have to be efficient and independent reduces the need for sophisticated types of responses and consequently limits the type of analytical tools that are needed to make sense of these responses. In ITS, tasks have to be ecologically valid and mimic real learning environments (for example, being conversational), and consecutive tasks are likely related. The response to the tasks can be multi-modal so advanced technologies have to be used to collect and interpret the

data. The ITS experience can thus help realise the vision of change for next-generation assessments described in previous chapters by illustrating ways to deliver intelligent feedback as well as providing models and tools to process complex, multi-modal data.

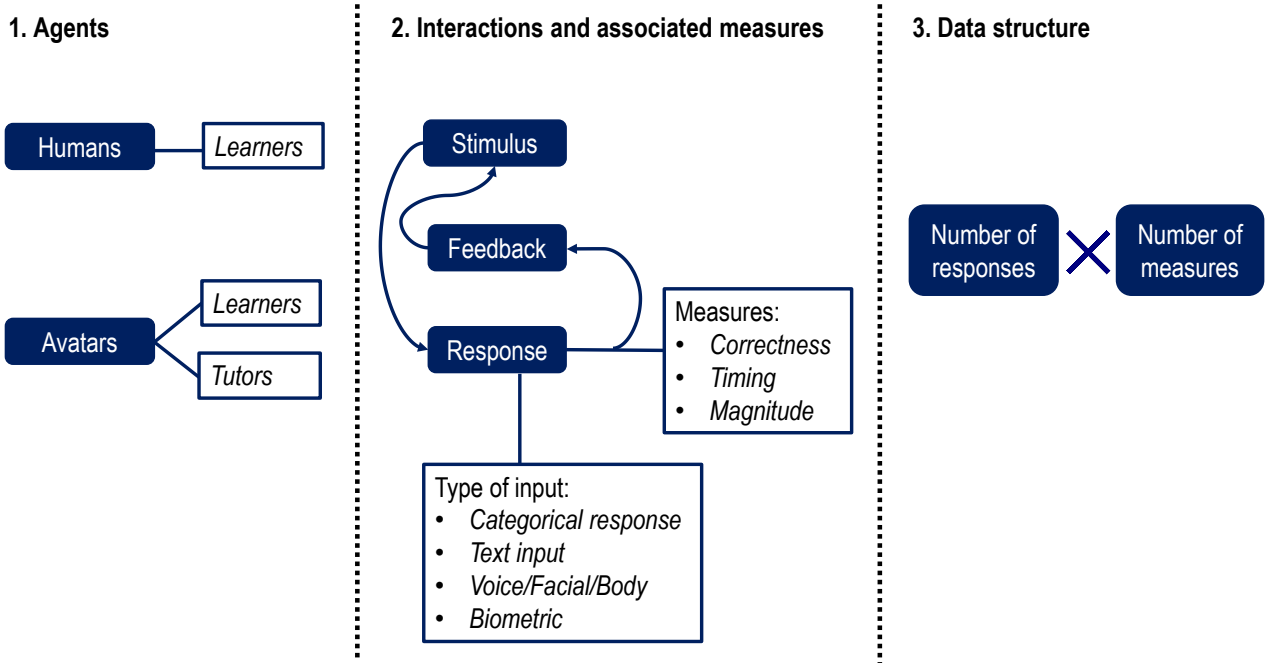
An ITS adaptive assessment framework

We propose a general framework for the use of ITS for assessment. The framework describes how ITS work, what types of data can be generated by the interactions between human learners and automated tutors, and how these data can be converted into evidence of students' knowledge and mastery of complex skills. The framework is an abstraction of the general elements characterising ITS that target different domains (e.g. from electricity to socio-emotional learning) and automatically process data emanating from different response processes (e.g. from multiple choice to open conversation and emotion detection). In particular, the framework shows how these very different sources of evidence can be organised according to the same data structure.

Figure 10.1 illustrates the main elements of the framework:

1. the type and function of agents that interact in the ITS.
2. the way agents interact through sequences of stimuli and responses.
3. the data structure corresponding to the number of responses multiplied by the number of measures that are calculated on each response.

Figure 10.1. Elements of an ITS adaptive assessment framework



The agents of Intelligent Tutoring Systems: Learners and avatars

In ITS, one or more human *learners* interact with one or more ITS applications called *avatars*. The avatars take either the role of a tutor, who explains concepts and asks questions, or of a peer learner, who

participates to the tutoring as the human learner does. Learners and avatars act similarly within ITS: both present questions for each other and offer answers with the best of their knowledge. The difference is that avatars are controlled by back-end AI-enabled engines that constantly assess learners during the interaction while the behaviours of human learners are controlled by back-end human intelligence (broadly defined) and are very likely doing similar assessments of the avatars.

The interactions between learners and avatars

The interactions between learners and avatars occur in a mixed-initiative style, where humans and avatars respond to questions of other humans and avatars. We characterise this iterative process as a series of sequences each involving a stimulus (S) and a response (R). Any action at the time *t* becomes the stimulus for another action at time *t+1*. For example, a typical interaction might start with the avatar tutor explaining a concept and then asking a question to learners to verify their understanding. This is the first stimulus at time *t*, that is followed by a response at time *t+1*. This response elicits feedback from another agent in the ITS at time *t+2* that results in a new stimulus. If the human agent provides an incorrect response, the avatar tutor might provide a clarification and invite the learner to give another response at time *t+3*. In this conversational environment, the number of responses that are collected from each human learner is thus not fixed but depends on the how the interaction unfolds between the human and the avatars.

In this framework, we distinguish two different aspects of the responses taking place in ITS: 1) the response type, or format of the response (e.g. text or voice input); and 2) the response measures, or types of evidence that are generated about student learning – including correctness as well as timing and magnitude. We discuss these two aspects in detail next.

Response types

Responses from learners in ITS can take one or a mixture of the following response types:

- Categorical behaviour that is easily identified, such as multiple choices.
- Written (typed) text responses.
- Natural language voice input.
- Facial expressions and body gestures.
- Bio-metrical response/behaviours captured through wearable devices, such as smart watches.

As shown in Table 10.1, collecting data from each response type requires different types of hardware in terms of assessment devices.

Table 10.1. Response types and input devices

Input (response type)	Hardware (device)	Software (processing)
Categorical response	Keyboard, mouse, touchpad	Classic statistics
Text input	Keyboard	Natural Language Processing (NLP)
Voice/facial/body	Voice recording, video camera	Speech-to-text, NLP, emotion detection, gesture detection
Physiological measures	Wearable devices	Big-Data analytical tools

Response measures

Evaluation of the correctness of responses

For each avatar action, the avatar has a set of "expected" correct responses and "speculated" wrong responses. Consider a simple and familiar case: if the action of the avatar is to present a multiple-choice question to the learner, the expected correct answer would be the correct choice and the speculated wrong responses would be the wrong choices. In this multiple-choice example, the avatar can always "score" the responses of any action into:

- *Hit (H)*: Correct choice is selected.
- *False Alarm (FA)*: Incorrect choice is selected.
- *Miss (M)*: Correct choice is not selected.
- *Correct Rejection (CR)*: Wrong choice is not selected.

Notice that when the response is correct, more than one score can be assigned: *Hit* (the correct selection) and *Correct Rejection* (the ones that are not selected). In the case in which more than one choice needs to be selected (such as "check all that apply" types), all four scores may exist in one response.

This simple categorical representation can be adapted to more complex responses from the human learner. No matter the type of response, within an ITS there is always a well-defined set of expected correct and incorrect responses. For each response type, the software processing the responses is responsible for classifying them into discrete responses:

- For categorical responses, such as multiple-choice responses, correct/incorrect responses are explicitly defined according to the scheme presented above (H, FA, M, CR).
- For non-categorical responses, such as text input, software is needed to classify the input text into letter grades or a more fine-grained evaluation against typical correct and incorrect answers.
- For voice input, software (some type of AI application) is needed to process and transcribe the input into natural language and then analyse it like text input from categorical responses.
- For facial or body gestures, software is needed to classify the input into discrete categories (such as emotions based on facial expressions).
- For biometric responses (collected from wearable devices, for example) data are larger than the previous types; specially designed software is needed to interpret student input.

Time-based measures

Given that the responses in ITS are recorded with timestamps and other physical information, it is relatively easy to measure the time course of each category and have the following measures that were not available for classical/traditional assessment instruments:

- *Response latency:* The time between the end of the previous stimulus and the starting of the response. This is one of the most studied dependent variables in cognitive psychology. It is easy to measure and very informative.
- *Duration:* The time between the action of the response and the submission of response. For example, if a learner is required to provide a drag-and-drop response, the duration would be time from the beginning of a dragging action to the time an object is dropped. When classifying facial expression into emotional states, some of the emotions may last longer than others.
- *Inter-category interval and intra-category interval:* When the processed result of a response includes multiple categories and the categories have different time courses, *inter-category* intervals are the difference between timestamps for different categories. *Intra-category* intervals consist of multiple categories. This measures the time between two observed categorical behaviours.

Measures of magnitude or intensity of the response

The technologies applied in modern ITS can also be used to classify responses according to other characteristics beyond their correctness and timing. One important additional measure is the *Magnitude*, which represents the intensity of the observed behaviour. A simple measure of magnitude can be obtained by asking the learner to rate their level of confidence in their response: if the learner states that he has high confidence in his response then the response is considered having high magnitude. Judgement of confidence can also be obtained unobtrusively with advanced software to process raw voice and obtain non-verbal information, such as hesitation, delay between sentences, intonations or sentiments. Measures of magnitude can have important applications in innovative assessments of social interaction skills and emotional regulation skills.

Figure 10.2 provides an illustration of the above measures for the simple scenario of a multiple-choice response type of task. In this example, the avatar asks a question and the human learner chooses among four responses options (A, B, C, D), where A, C, D are correct choices and B is the wrong choice. In the first response, the learner indicates option A and B as correct: there is thus one Hit (A), two Misses (C and D) and one False Alarm (B). The avatar reacts at *t+1* providing feedback and then the learner changes his response, indicating A and C as correct. There is an improvement from 1 out of 4 correct categories to 3 out of 4 with respect to the first response (the only issue in the second response is the missing D). After receiving feedback from the avatar again the learner submits a third response, this time correctly selecting options A, C and D and rejecting B. In Figure 10.2, the height of the bar represents the *magnitude* while the width represents the length of time between the stimulus and the submission of the responses (*response latency*). In this representation, an *inter-category interval* is given by the time the student takes to shift from an incorrect to a correct response (or vice versa), while an *intra-category interval* is given by the time the student keeps a response the same.

Figure 10.2. The interaction between learners and avatars (tutors) in ITS

Learner response categories for assessments that require categorical responses with answers and distractors

Notes: A, B, C, D represent the response options that the learner can select in a multiple-choice task. Grey vertical lines represent the different iterations between the learner and the tutoring system, from the start of the task to the different responses from the learner and feedback from the tutor. Responses are processed following the Hit (H), False Alarm (FA), Correct Rejection (CR), Miss (M) categories, and green and red colours flag correct and incorrect learner responses respectively. In each response, the height of the bars represents the magnitude, while the width represents the length of time between the stimulus and the submission of the responses (response latency).

Data structure

If the number of independent behavioural categorical measurements for each response is *N*, the data for adaptive assessments in ITS will form a *K* x *N* data matrix, where *K* would be the total number of responses. Take, for example, the data for Figure 10.2. There are four behaviour categories: *Hit* (H), *False Alarm* (FA), *Miss* (M), and *Correct Rejection* (CR). If the tutor only provides two feedback interventions, then the data matrix would be a 3 by 4. If additional measures are collected (e.g. confidence and timelapse of the responses), there will be additional values in each cell (e.g. response latency, duration, and magnitude). The other two measures (inter-category interval and intra-category interval) and second-order numerical properties of the data matrix can be computed easily.

This general data structure is applicable to a variety of adaptive test environments. The next section covers a few examples demonstrating the utility of the framework.

Selected examples

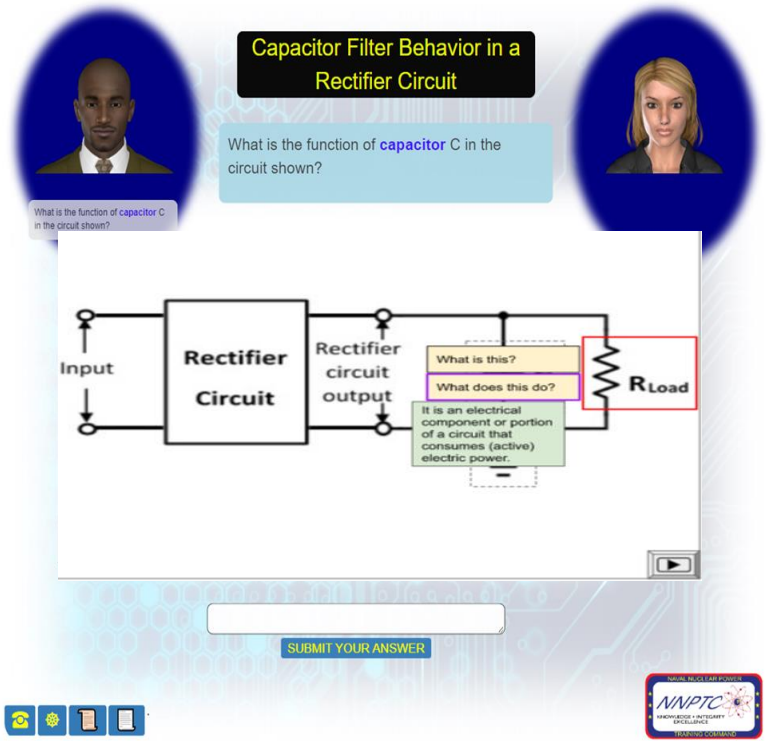
In the previous section, we presented a framework for adaptive assessments in ITS, which was illustrated with clearly defined response categories (H, FA, M, CR). In the examples we present next, the assessment framework is the same, but response categories vary. The first example looks at conversation-based ITS by focusing on two applications of AutoTutor. The second example is an application of the framework in assessing competencies of team members in group interaction. In the third example, we present an application of the framework to an assessment of emotional responses.

Example 1: Conversation-based ITS

In this first example, we look at two applications of AutoTutor. The first one, ElectronixTutor (Graesser et al., 2018[10]; Morgan et al., 2018[11]), is an ITS including 30 modules each covering a key concept in one of five areas of electronics (semiconductor, PN junction, rectifiers, filters and power supplies). In ElectronixTutor, learners interact with the tutor by typing (or voice input) in plain English. As shown in Figure 10.3, the tutor starts with a seed question about the concept (e.g. “What are the disadvantages of a bridge rectifier?”). In turn, using a combination of latent semantic analysis (LSA) and RegEx, the ITS evaluates learner responses by analysing their semantic similarity to typical expectations or typical misconceptions – that is, matching them against a set of previously defined correct and incorrect answers (see Figure 10.4). If the learner input is incorrect, the tutor provides a hint for the learner to respond to; if the response to the hint is incorrect, the tutor provides a narrower hint targeting a specific word or phrase in the form of a “prompt”. Finally, the tutor provides an assertion or summary of the expectation and moves on to providing hints, prompts and assertions for the next expectation.

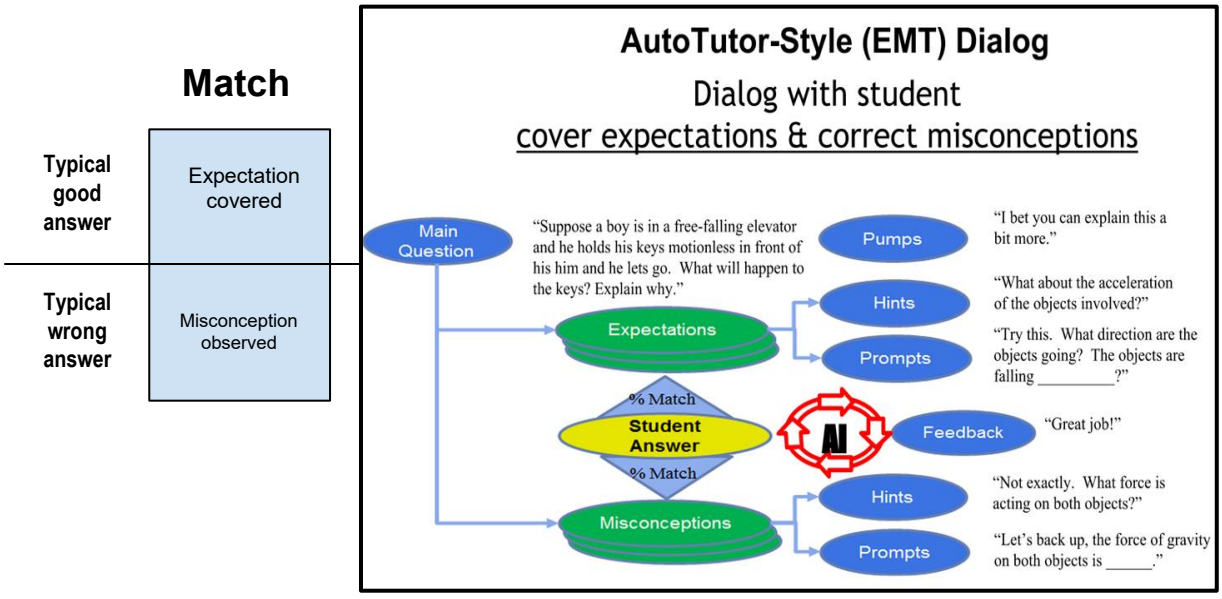
It must be noted that, while we considered four indicators of correctness for learner responses in the previous section (*Hit*, *Miss*, *False Alarm*, *Correct Rejection*), ElectronixTutor considers two: *Hits* (i.e. responses that meet the semantic overlap threshold to the ideal answer) and *Misses* (i.e. those that do not meet the semantic overlap threshold). The tutor provides just-in-time feedback for misses and positive feedback for hits. All interaction data is stored in a learning record store and can be drawn upon to display learners’ performance, both to instructors and learners themselves through a dashboard user interface. The dashboard can be easily integrated into most learning management systems meaning that the modules of the application can be used for learning and assessment in the classroom. ElectronixTutor is designed to offer two types of feedback (hints and prompts) for both typical expectations and misconceptions. The data structure in this example is hence a 2 by 2 matrix.

Figure 10.3. ElectronixTutor interface



Source: Screenshot of ElectronixTutor user interface.

Figure 10.4. The conversation flow of an Expectation-Misconception Tailored dialogue

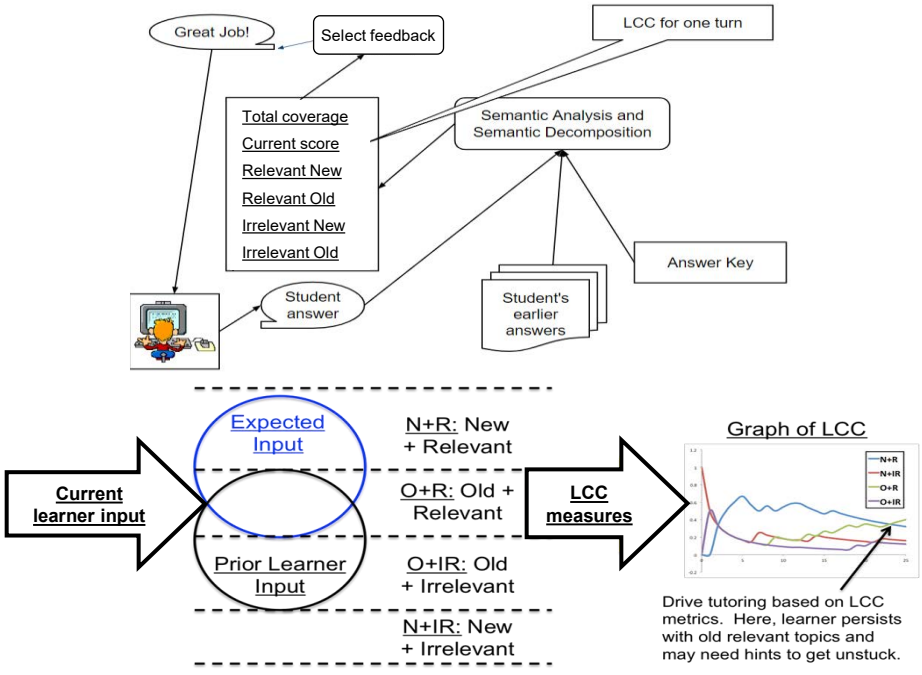


A different, simpler version of AutoTutor, LITE (Hu et al., 2009[12]), provides a model where students are presented with a deep reasoning question about a concept that requires explanation beyond a "yes" or "no" response (Hu et al., 2009[12]; Sullins, Craig and Hu, 2015[13]). In this variation, once the student has provided an answer, a language analyser semantically "decomposes" it (Hu et al., 2014[14]) and computes its relevance against the answer key (i.e. correct answer). The numerical value (between 0 and 1) of semantic similarity between the student's answer and the answer key is used as the measure of *relevance* (R) and *irrelevance* (IR). The numerical value of irrelevance is 1 minus the numerical value of relevance. If the student does not give a satisfactory answer, the tutor will invite the student to provide further input. The analyser will then decompose the new response to determine, first, whether it overlaps semantically with the previous one or not – this is, to see if the student is providing new (N) or old (O) information – and second, whether the information is relevant or not. With this process, a sequence of vectors can be produced, each vector containing six values: *Relevant-New* (N+R), *Relevant-Old* (O+R), *Irrelevant-New* (N+IR), and *Irrelevant-Old* (O+IR), in addition to *Total Coverage* (accumulated pure relevant and new (N+R) from the first answer to the last answer) and *Current Score* (most recent (N+R) + (O+R)). All six scores, not necessarily independent, provide enough information for the ITS to give feedback.

Computation of these values can be achieved by most current semantic analytical tools (Hu et al., 2014[14]). Each element of the vector will serve as an assessment for the students' understanding of the given concept. We call the sequences of values (O+R, O+IR, N+R and N+IR) a learner's characteristic curve (see Figure 10.5), which has a rather intuitive interpretation that can be easily used to derive feedback:

- When the student does not know the answer or does not know what was asked, the values of (N+IR) are likely high and this indicates confusion.
- When the student misunderstands the question or has misconception, the values of (O+IR) are likely high.
- High values of (O+R) indicate that the student may have answered with all the knowledge.
- High values of (N+R) indicate that the student is likely to contribute more towards the answers.

Figure 10.5. Learner's characteristic curve



Source: Hu et al. (2013[15]).

Example 2: Assessing competencies of team members in group interaction

The next example demonstrates that the framework presented here can be used to assess learners' competencies when interacting with others. To illustrate this possibility, imagine that learners and avatars in an ITS contribute to a discussion by responding to the previous contributions from other learners or avatars. Assume the responses are finite and happen at discrete time points ($t_1,...,t_n$), where each action at time t_k is the response to some or all of actions prior to k. There are no assigned actions meaning that each learner/avatar decides to act and some of them may act more than others.

In such context, different types of evidence are produced as learners/avatars respond to previous actions at any time t. As summarised in Table 10.2 below, such evidence can be used to calculate six indices of a group communication analysis (GCA) vector for each learner/avatar (Dowell, Nixon and Graesser, 2018[16]; Hu et al., 2018[17]).

Table 10.2. Indices of group communication derived from interactions in an ITS generic scenario

Indices	Definition	Method of computation
Participation	Measures how much a learner participates in the group. At each occurrence, this value will increase for the learner but will decrease for all other learners and avatars, since they are not acting at this time.	Cumulative frequency of contributions.
Overall responsivity	The value will increase if the action of the learner is actively in response to previous actions.	Semantic similarity of a contribution in relation to the contributions of other participants.
Internal cohesion	This value will increase if the action is consistent to the previous actions of the learner.	Semantic similarity between current contribution and previous contributions of the same learner.
Social impact	This value is not updated for the current learner. However, values for all other learners/avatars may change depending on how much the contribution of the current learners is related to their perspective actions.	Semantic similarity between one agent's past contributions and current contributions of others.
Newness	This value will increase if the action of the learner is different from its previous actions.	Semantic distinctiveness between current contribution and the same agent's previous contributions.
Communication density	Measured by how meaningful the contribution is.	Semantic density of the contributions (i.e. it can be measured by domain-relevant key terms).

Note: Detailed computation for the indices can be found in Table 1 in Hu et al. (2018[17]).

To compute these indices, behaviour of the participants needs to be recorded – for example, the details of each contribution of the participants, such as the time, content (language) and the target (who he/she is addressing to). Some advanced linguistic analytical tools, such as semantic analysis (Hu et al., 2014[14]; Hu, Cai and Olney, 2019[18]), are needed for some of the indices.

With these indices computed after each round of contributions, a sequence of 6-dimensional GCA vectors can be obtained. The elements of the 6-dimensional vector are measures of response categories. With this sequence of vectors, timestamp (latency, duration) can be associated with some of the values. There are also "expected" correct actions and "speculated" wrong moves associated with the actions, so similar classifications can be performed.

In this example, the data structure is K x 6, where K is the number of contributions from an agent. Notice that even if each of the contributions is only made by one individual, the GCA vectors are updated for all participants. Additionally, although the data structure is the same as in the previous example, the nature of the behavioural categories is different. The potential application for this example is to assess "teamness" of individuals taking part in artificially constructed discussion environments, where some or all other participants are avatars controlled by ITS.

Example 3: Assessment of emotional responses

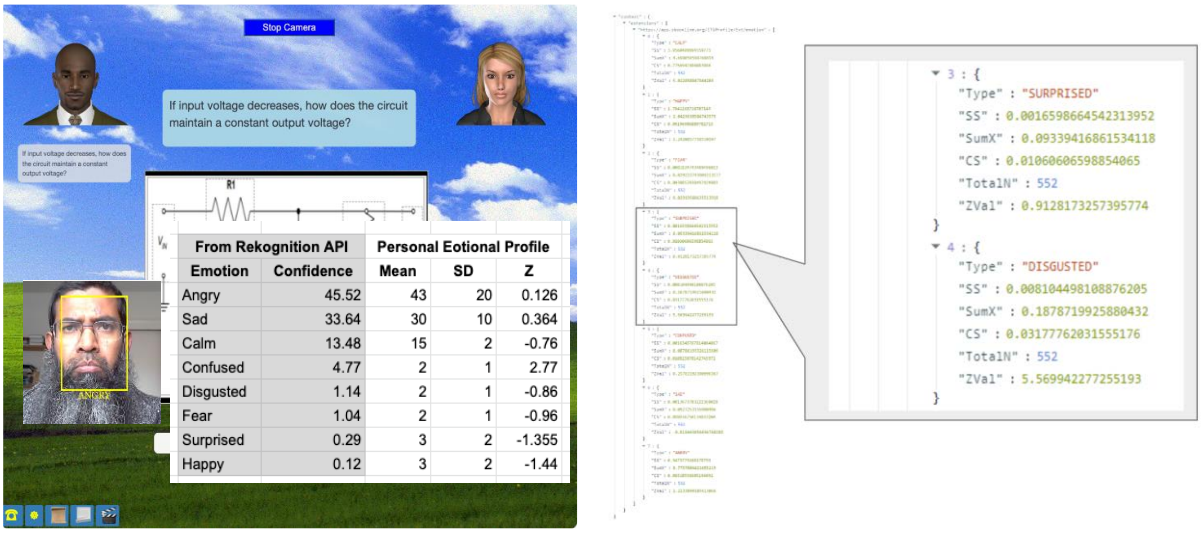
Emotions and cognitive affective states affect learning in different ways. For example, meta-analyses of emotions during learning suggest that emotions that frequently occur include boredom, frustration, confusion, happiness, anxiety and flow/engagement (D'Mello and Graesser, 2012[19]). Previous ITS, to varying degrees of success, have implemented affect-detection measures meant to enhance feedback and interventions during the tutoring process (D'Mello, Picard and Graesser, 2007[20]; D'Mello and Graesser, 2012[19]). But how can information about a learner's affective state improve adaptive assessments?

One possibility is to use facial recognition software – for example, the sensory model of the Generalised Intelligent Framework for Tutoring (GIFT) (Sottilare et al., 2017[3]) – to capture learning-relevant affective states (Ahmed, Shubeck and Hu, 2021[21]). Updating the student model with information on their affective states can potentially help assess a learner's knowledge state, which goes beyond measures of past performance or simple self-reported measures of confidence. For example, learners that appear confused when responding to a question may indicate that they guessed the answer. Similarly, learners that appear to be frustrated may indicate low confidence on a certain topic or knowledge component. Learners that appear to be happy or in a state of flow (i.e. a cognitive state where one is completely immersed in an activity) may suggest high confidence and understanding of the material. However, if a learner appears confident when providing an incorrect response this may suggest an existing misconception that should be addressed.

In a traditional non-adaptive assessment environment, it can be difficult to determine if an incorrect response was due to carelessness, genuine lack of understanding or simply not attending to the task. Within the context of the *Hit*, *False Alarm*, *Miss* and *Correct Rejection* framework, knowing a learner's affective state at each interaction point can provide further insight into why a learner changed their response from a *False Alarm* to a *Hit* or from a *Correct Rejection* to a *False Alarm*. Consider, for example, a learner who appears confused when providing an incorrect response to a question and then later appears calm or happy when making a correct response for their second attempt. This could suggest that the confusion was resolved. Here, the affective data is used to enhance the system's confidence in the learner's current knowledge state. Mapping a learner's affective state onto relevant knowledge components can strengthen the system's student model. This improved student model can then be used to provide data-driven, "learner aware" (Ahmed et al., 2020[22]) and affect-sensitive interventions and feedback to potentially guide the user back into an affective state conducive to learning.

In a pilot implementation, a simple Amazon Rekognition API (AWS, 2023[23]) was integrated in an existing ITS (see Figure 10.6). Facial expressions of learners were processed every 500 milliseconds and mapped each time onto eight emotional states (Angry, Sad, Calm, Confused, Disgusted, Fear, Surprised and Happy) with varying degrees of confidence – for example, Rekognition API classified the face as Angry with 45.52% confidence. It must be noted that the AI behind the Rekognition API is based on a large number of human faces, but it may not be accurate when classifying emotions for individuals in certain contexts, such as learning. It may be, for instance, that the facial expression of the individual in Figure 10.6 can be generally classified as Angry (with 45% confidence from API) but, for this particular individual, it is a typical expression of confusion when working with difficult tasks (a z-score of 2.77 for the emotion type of confusion). Indeed, the outcome of the eight emotion types from Rekognition API may not be the perfect measure of emotion types during learning. However, having a numerical vector that are sensitive to learners' emotional change as function of tasks during learning made it possible for potential applications for assessment.

Figure 10.6. Measuring emotional responses during tutoring



Note: Facial emotional data are sampled in fixed intervals (500 milliseconds). An eight-dimension emotional vector is computed from Rekognition API and then z-scores for each of the dimensions are computed from the distributions of measures.
Source: Combined screenshots from Amazon Rekognition API (AWS, 2023[23]).

Conclusion and final thoughts

In this chapter, we introduced an adaptive assessment framework inspired by Intelligent Tutoring Systems. In this framework, regardless of the chosen domain and type of task, the data structure is in the form of *K* x *N*, where *K* is the number of responses and *N* is the number of assessment categories. The adaptive nature of assessments in this framework, however, requires models that analyse sequences of actions in interconnected tasks. The examples presented in this chapter have shown that the current state of AI makes it possible to interpret data and develop indicators for sequences of actions with multiple data types – albeit several limitations remain, both theoretical and technological. Theoretically, we still need to understand what exactly the fine distinctions between learning environments and assessment environments are. What we have presented is based on a typical assumption in a learning framework where any exposure of learning resources (even in the form of assessment items) will have an explicit or implicit impact on future performance. Yet, because this process is adaptive and dynamic, it is necessarily different for each learner. In other words, the assessment of a learner's ability will likely be process dependent. It would be a challenging task to design an *adaptive* assessment that measures the same competence for all the learners.

There are a few technological issues as well. The technology that we use today may be outdated tomorrow. For example, there are new and improved semantic processing technologies today that are much better than those used five years ago. Different technology will necessarily produce different adaptive processes and there will be issues if longitudinal comparisons are made using different technologies at different times. In addition, there are large differences in terms of the availability of technologies across different countries and regions. For example, semantic analytic tools are more mature for some languages, such as English, than for others. Using technologically dependent process data to assess students' ability will need to be validated when the availability of technologies changes. The processing of natural language input, such as through syntactic parsing or semantic encoding, is more computationally expensive (both as the user terminal and server) than processing categorical inputs. Differences in the availability of technology and computational power might result in differences in access to innovative systems across countries.

Furthermore, while AI-powered learning environments offer new avenues to analyse the process and not only the product of learning (e.g. through process data), and to emulate real-life, open and dynamic contexts (e.g. by introducing complex conversations), their ability to enhance the authenticity of assessments remains limited in some respects (Swiecki et al., 2022[24]). For instance, despite the progress of natural language processing, developing AI systems capable of processing the full spectrum of human expression including elements such as humour or double meanings remains elusive. Limited "reasoning" abilities hold back AI's capacity to categorise learner behaviour well in truly open environments where qualitative information is needed, compromising in turn the offering of "intelligent" tutoring feedback. Further limitations include increasingly recognised issues of discrimination stemming from the fact that AI systems build on models, both expert-based and data-driven, grounded on restricted ideas of learning and behaviour pertaining to particular groups of learners (e.g. white, male, neurotypical students). AI technologies must become more mature if we are to rely on non-human intelligences for the provision of the authentic and adaptive tutoring and assessment experiences that all learners deserve.

References

Ahmed, F. et al. (2020), "Enable 3A in AIS", in *HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games, Lecture Notes in Computer Science*, Springer, Cham, https://doi.org/10.1007/978-3-030-60128-7_38. [22]

Ahmed, F., K. Shubeck and X. Hu (2021), "Enhancement of GIFT Enabled 3A Learning: New additions", *Proceedings of the 9th Annual Generalized Intelligent Framework for Tutoring (GIFT) Users Symposium (GIFTSym9)*. [21]

Avvisati, F. and F. Borgonovi (2020), "Learning mathematics problem solving through test practice: A randomized field experiment on a global scale", *Educational Psychology Review*, Vol. 32/3, pp. 791-814, https://doi.org/10.1007/s10648-020-09520-6. [4]

AWS (2023), *Amazon Rekognition: Developer Guide*, Amazon Web Services. [23]

Butler, A. (2010), "Repeated testing produces superior transfer of learning relative to repeated studying", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 36/5, pp. 1118-1133, https://doi.org/10.1037/a0019902. [5]

Dempster, F. (1996), "Distributing and managing the conditions of encoding and practice", in Ligon Bjork, E. and R. Bjork (eds.), *Memory*, Elsevier, https://doi.org/10.1016/b978-012102570-0/50011-2. [6]

D'Mello, S. and A. Graesser (2012), "AutoTutor and affective autotutor", *ACM Transactions on Interactive Intelligent Systems*, Vol. 2/4, pp. 1-39, https://doi.org/10.1145/2395123.2395128. [19]

D'Mello, S., R. Picard and A. Graesser (2007), "Toward an affect-sensitive AutoTutor", *IEEE Intelligent Systems*, Vol. 22/4, pp. 53-61, https://doi.org/10.1109/MIS.2007.79. [20]

Dowell, N., T. Nixon and A. Graesser (2018), "Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions", *Behavior Research Methods*, Vol. 51/3, pp. 1007-1041, https://doi.org/10.3758/s13428-018-1102-z. [16]

Graesser, A. et al. (2018), "ElectronixTutor: An intelligent tutoring system with multiple learning resources for electronics", *International Journal of STEM Education*, Vol. 5/1, https://doi.org/10.1186/s40594-018-0110-y. [10]

Hu, X. et al. (2009), "AutoTutor lite", *Artificial Intelligence in Education*. [12]

Hu, X., Z. Cai and A. Olney (2019), "Semantic representation and analysis (SRA) and its application in conversation-based Intelligent Tutoring Systems (CbITS)", in Feldman, R. (ed.), *Learning Science: Theory, Research, and Practice*, McGraw-Hill Education. [18]

Hu, X. et al. (2018), "Constructing Individual Conversation Characteristics Curves (ICCC) for Interactive Intelligent Tutoring Environments (IITE)", in Sottilare, R. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems*, US Army Research Laboratory, Orlando. [17]

Hu, X., D. Morrison and Z. Cai (2013), "Conversation-based intelligent tutoring system", in Sottilare, R. et al. (eds.), *Design Recommendations for Intelligent Tutoring Systems: Learner Modeling*, U.S. Army Research Laboratory, Orlando. [15]

Hu, X. et al. (2014), "Semantic representation analysis: A general framework for individualized, domain-specific and context-sensitive semantic processing", in Schmorrow, D. and C. Fidopiastis (eds.), *Foundations of Augmented Cognition. Advancing Human Performance and Decision-Making through Adaptive Systems, Lecture Notes in Computer Science*, Springer, Cham, https://doi.org/10.1007/978-3-319-07527-3_4. [14]

Morgan, B. et al. (2018), "ElectronixTutor integrates multiple learning resources to teach electronics on the web", *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, https://doi.org/10.1145/3231644.3231691. [11]

Nwana, H. (1990), "Intelligent tutoring systems: An overview", *Artificial Intelligence Review*, Vol. 4/4, pp. 251-277, https://doi.org/10.1007/bf00168958. [1]

Roediger, H. and J. Karpicke (2006), "The power of testing memory: Basic research and implications for educational practice", *Perspectives on Psychological Science*, Vol. 1/3, pp. 181-210, https://doi.org/10.1111/j.1745-6916.2006.00012.x. [7]

Rowland, C. (2014), "The effect of testing versus restudy on retention: A meta-analytic review of the testing effect", *Psychological Bulletin*, Vol. 140/6, pp. 1432-1463, https://doi.org/10.1037/a0037559. [8]

Rus, V. et al. (2013), "Recent advances in conversational Intelligent Tutoring Systems", *AI Magazine*, Vol. 34/3, pp. 42-54, https://doi.org/10.1609/aimag.v34i3.2485. [2]

Sottilare, R. et al. (2017), "The Generalized Intelligent Framework for Tutoring (GIFT)", in Galanis, G., C. Best and R. Sottilare (eds.), *Fundamental Issues in Defense Training and Simulation*, CRC Press, London, https://doi.org/10.1201/9781315583655-20. [3]

Sullins, J., S. Craig and X. Hu (2015), "Exploring the effectiveness of a novel feedback mechanism within an intelligent tutoring system", *International Journal of Learning Technology*, Vol. 10/3, p. 220, https://doi.org/10.1504/ijlt.2015.072358. [13]

Swiecki, Z. et al. (2022), "Assessment in the age of artificial intelligence", *Computers and Education: Artificial Intelligence*, Vol. 3, pp. 1-10, https://doi.org/10.1016/j.caeai.2022.100075. [24]

Wheeler, M. and H. Roediger (1992), "Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results", *Psychological Science*, Vol. 3/4, pp. 240-246, https://doi.org/10.1111/j.1467-9280.1992.tb00036.x. [9]

Part III Innovating How We Interpret and Use Assessment Results

11 Cross-cultural validity and comparability in assessments of complex constructs

By Kadriye Ercikan, Han-Hui Por and Hongwen Guo

(Educational Testing Service)

This chapter discusses issues and methodologies in cross-cultural validity and comparability in multilingual and multicultural assessment contexts, considering construct equivalence, test equivalence and testing condition equivalence. It discusses the important role of sociocultural context in shaping learning and assessment performance, particularly for assessments measuring 21st Century competencies like creativity or critical thinking. Cultural norms, practices and opportunities to learn affect students' understanding of assessment items, problem-solving strategies and performance. In innovative digital assessment contexts, students' digital literacy and potential biases in test adaptivity and artificial intelligence-based automated scoring and item generation engines also necessitate consideration. The chapter highlights recommendations for evaluating and optimising validity and comparability in innovative assessments of 21st Century competencies.

Introduction

Assessments of 21st Century competencies, such as complex problem solving, are expected to be engaging, resemble real life tasks, draw upon multidisciplinary knowledge and skills, and provide individuals with feedback on their progress towards solving problems. Several testing innovations are used in large-scale digital assessments to meet these goals (see Chapters 5 and 7 of this report). These innovations include adaptivity based on performance on segments of the assessment, interactivity that modifies the assessment as determined by student actions during test-taking, and the use of multimedia tools and digital features of assessment environments. Interactivity and adaptivity in assessments, as well as the need for producing more engaging assessment, immediate and effective feedback to learners, and cost efficiency, can also utilise artificial intelligence (AI)-based tools such as automated item generation and AI-based automated scoring. While these innovations enhance and help meet the demands for such assessments, we must consider the new sources of threats to cross-cultural validity and comparability of assessment inferences that they generate. These implications are discussed in this chapter.

Throughout the chapter, we use *assessment* to describe educational and psychological measurement instruments, including surveys and questionnaires, and related documents and procedures such as instructions, scoring guidelines and administration procedures. The term *test* refers specifically to a test form or when it is part of a commonly used phrase, such as in "test equivalence" or "test adaptation". In describing the mode of assessments, we use the terms *digital* or *innovative* to refer to the broad classes of technology-based assessments (TBAs) and technology-enhanced assessments (TEAs). Furthermore, *measurement/score comparability* and *measurement equivalence* are used interchangeably and refer to the comparability of score interpretation and use and the statistical notions of measurement equivalence.

Sociocultural context of assessment

Students' participation in assessments has been viewed typically through a cognitive perspective which focuses on teaching, learning and performance on assessments without taking the sociocultural context into account (Pellegrino, Chudowsky and Glaser, 2001[1]). This perspective is in contrast to the sociocultural lens, also referred to as the "situated perspective" (Gee, 2008[2]), which considers learning and performance on assessment in terms of the relationship between individuals and the social environment and the context in which they think, feel, act and interact (Moss et al., 2008[3]).

There is extensive research evidence that social and cultural contexts can affect learning and world views, including how success is perceived, how students are taught and how achievement is defined in education systems. For example, Yup'ik children in rural Alaska learn critical community practices such as fishing and navigation from observing and participating in these activities with experienced adults. Because verbal interactions are part of this key learning process, a school system that expects passive listening with little contextual interaction may disadvantage these students (Lipka and McCarty, 1994[4]).

Sociocultural context plays a particularly critical role in the development of complex 21st Century constructs such as creativity, critical thinking, problem solving or collaborative skills. These constructs are defined in terms of how students think, feel, empathise, act and interact with others and their social environment, and are grounded in social and cultural contexts (Suzuki and Ponterotto, 2007[5]). Research points to evidence of differences in the conceptual definitions and applicability of such constructs in different cultures (Ercikan and Lyons-Thomas, 2013[6]; Ercikan and Oliveri, 2016[7]; Niu and Sternberg, 2001[8]; Suzuki and Ponterotto, 2007[5]; Lubart, 1990[9]).

The central role that social and cultural context plays on learning extends to assessments. When students engage with assessments, their experiences, language and sociocultural backgrounds interact with the knowledge and skills targeted by the assessment, which can in turn impact their performances (Liu, Wu and Zumbo, 2006[10]; Solano-Flores and Nelson-Barber, 2001[11]). Responses to test items reflect what

students know and can do as well as a complex interaction of how they feel about the assessment situation, understand assessment questions and formulate their responses that is based on their social environment and cultural and language practices outside of school.

Cultural norms and practices

The nature and frequency of access to cultural practices outside of school can affect students' understanding of assessment items and hence their performance. For example, students who attend French schools as linguistic minority students in Ontario, Canada, where English is the dominant societal language, consistently performed worse in mathematics, reading and science than French students in Quebec, Canada, where French is the dominant language (Ercikan et al., 2014[12]). Similar performance trends have been observed for other ethnic, racial and linguistic minority groups in other countries (Ercikan et al., 2014[12]; Ercikan and Elliott, 2015[13]). In each occurrence there are likely multiple factors contributing to these patterns, but the contribution of students' social and cultural contexts as well as access to the language of schooling need to be considered as important factors for learning and assessment outcomes.

Parental involvement and expectations – another contributor to the social aspect of learning – have also been found to play a role in academic performance (Cooper et al., 2009[14]; Lee and Stankov, 2018[15]). For example, students in Asian countries, such as China, view their education and examination systems as the main route out of poverty. This context, combined with being praised and rewarded for good performance by teachers and parents, is associated with higher motivation to perform well on the assessment (Rotberg, 2006[16]; Zhang and Luo, 2020[17]). Students from Asian cultural groups tend to have higher engagement levels with low-stakes assessments than students from some Western countries (Ercikan, Guo and He, 2020[18]; Guo and Ercikan, 2021[19]), also pointing to higher motivation levels.

Opportunity to learn and assessment performance

An important context that can affect students' assessment performance is the opportunities available to study a topic, learn how to solve the type of problems included on an assessment, and engage with similar kinds of assessments (Ercikan, Roth and Asil, 2015[20]; National Academies of Sciences, Engineering, and Medicine, 2019[21]). The opportunity to access the curricular content which is subsequently assessed, develop test-taking strategies and become familiar with a given assessment technology can all contribute to students' ability to engage with an assessment.

The role of *opportunity to learn* (OTL) in performance on assessment and in the validity of interpretation and use of assessment results has been widely recognised (Moss et al., 2008[3]; National Academies of Sciences, Engineering, and Medicine, 2019[21]). Large-scale assessments like the National Assessment of Educational Progress (NAEP) and the Programme for International Student Assessment (PISA) include measures of OTL to facilitate better interpretation and use of performance results (National Center for Educational Statistics, 2021[22]; OECD, 2016[23]; Schleicher, 2019[24]). Indicators of OTL often include school poverty rates, access to high-quality preparatory (e.g. kindergarten) programmes and coursework, access to high-quality teaching, curricular breadth and academic support (National Academies of Sciences, Engineering, and Medicine, 2019[21]). As OTL tends to measure access to curricular programmes or academic support, many of the common OTL indicators have demonstrated association with academic performance such as applied mathematics problems (Schmidt, Zoido and Cogan, 2014[25]) and advanced mathematical concepts (Cogan, Schmidt and Wiley, 2001[26]; Fuchs and Wößmann, 2007[27]; Schmidt et al., 2015[28]).

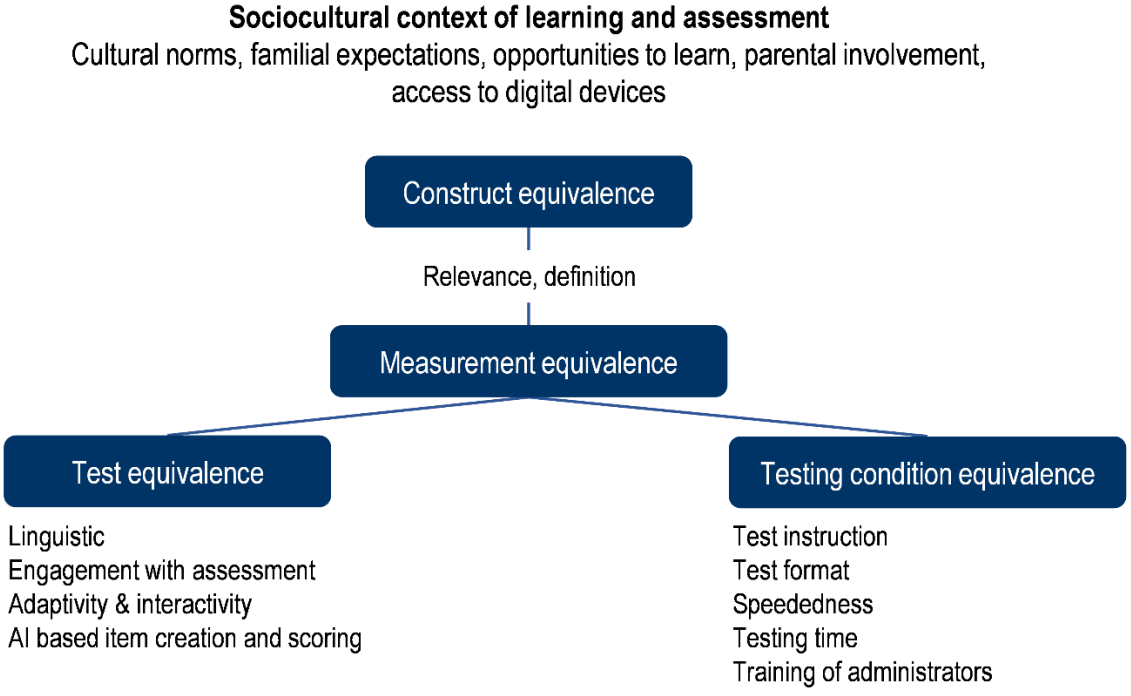
While many of these indicators that primarily capture opportunities to learn in school have obvious and shown associations with learning outcomes, their relationship with competencies like problem solving, decision making or collaborative skills (Kurlaender and Yun, 2005[29]; 2007[30]) are less definitive as the development of such constructs are not as closely connected to specific subjects, concepts and

procedures that are taught at school. Therefore, OTL indicators for these types of constructs may need to include those opportunities outside of schools, such as in the home and in broader societal contexts. Additionally, while within-group differences in OTL create challenges for score interpretation at the group levels, measuring the OTL of a complex construct remains a valuable exercise. For example, variations within and across cultural groups in opportunities to learn and practice problem solving can provide important insights to guide policy and practice in reducing such disparities.

Considerations for cross-cultural validity and comparability in assessments

Sociocultural context plays an important role in shaping learning and performance on assessment, which therefore has clear implications for assessment validity (and consequently for establishing validity evidence to support measurement equivalence). *Validity* is defined as the "degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, NCME, 2014, p. 11[31]). Similarly, *cross-cultural validity* refers to the degree to which evidence and theory support the interpretations and uses of test scores for different cultural groups and comparisons across groups. 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.

Figure 11.1. Issues affecting measurement equivalence across cultural and linguistic groups



As described in Figure 11.1, measurement equivalence requires evidence that tests are capturing equivalent constructs (*construct equivalence*) across subgroups, have similar measurement characteristics and properties (*test equivalence*), and they are administered in equivalent conditions (*equivalence of testing conditions*) (Ercikan and Lyons-Thomas, 2013[6]; Perie, 2020[32]). All three aspects of measurement equivalence are inherently influenced by sociocultural context.

Construct equivalence

The equivalence of constructs is the degree to which construct definitions are similar for populations targeted by the assessment, whether individuals are expected to develop and progress on these constructs in similar ways, and whether they 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[7]). 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[9]; Niu and Sternberg, 2001[8]). 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[35]).

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 of student outputs to vary across cultures. Because of these differences across cultural groups, there is a need to balance measurement validity with score comparability (see Box 11.1 for an example in the context of a large-scale assessment).

Box 11.1. Optimising comparability in the PISA 2022 Creative Thinking assessment

The PISA 2022 assessment of creative thinking exemplifies how an assessment of a complex construct across language and cultural groups can focus on certain aspects that optimise comparability. The developers of the assessment emphasised that the items should draw upon "the knowledge and experiences that are common to most students around the world" and reflect "the types of creative expressions that 15-year-old students can achieve and that can be most meaningfully and feasibly assessed" within the constraints of a PISA environment (Foster and Schleicher, 2021[33]).

To ensure this, the assessment developers addressed five issues in particular. These included:

- Focusing the assessment on the narrower construct of creative thinking, which was defined as "the competence to engage productively in the generation, evaluation and improvement of ideas" (OECD, 2022[34]). This narrower focus emphasised the cognitive processes related to idea generation whereas the broader construct of creativity also encompasses personality traits and requires more subjective judgements about the creative value of students' responses.
- Defining creative thinking, how it is enabled (i.e. OTL indicators of creative thinking) and what it looks like in the context of 15-year-olds in the classroom, focusing on aspects of the construct that are more likely to be developed in schooling contexts rather than outside of school.
- Identifying cross-culturally relevant assessment domains in which 15-year-olds would be able to 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, etc.).
- Focusing 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 bias).
- Engaging in significant cross-cultural verification of the coding rubrics, including using sample responses from students in several countries.

Test equivalence

Test equivalence requires equivalence of test versions in different languages and for different cultural groups in multilingual assessment contexts. In innovative, performance-based assessments, another key aspect of test equivalence is the degree to which students, including those from different language and cultural backgrounds, perceive and engage with the tasks in the same way. Test designs – especially those featuring adaptivity or AI tools in item creation and scoring – have direct implications on test content and scores, meaning they are also important to consider for test equivalence. In this section, we review various factors related to test equivalence including linguistic comparability, equivalence of test engagement, adaptivity and AI-based item creation and test scoring.

Linguistic comparability of assessments

In examining cross-cultural validity and comparability, particular attention has been given to the comparability of different language versions of assessments. In multilingual assessment contexts, test adaptation facilitates the administration of assessments to students in their language of instruction to provide a valid measurement of the targeted constructs.

The task of test adaptation goes beyond the literal translation of the assessment content (Ercikan and Por, 2020[36]; Perie, 2020[32]; van de Vijver and Tanzer, 1997[37]). An abundance of research has shown that a literal translation does not necessarily produce equivalent measurement and therefore comparability of scores due to key differences between languages (Allalouf, Hambleton and Sireci, 1999[38]; El Masri, Baird and Graesser, 2016[39]; Ercikan and Koh, 2005[40]; Hambleton, Merenda and Spielberger, 2005[41]). Languages vary in the frequency of compound words, word length, sentence length and information density (Bergqvist, Theens and Österholm, 2018[42]). Moreover, grammatical forms in one language may not have equivalent forms in other languages or may have many of them. There is also the difficulty of adapting syntactical style from one language to another, and languages may also differ in form (alphabet versus character-based) and direction of scribe (left-to-right, right-to-left or top-to-bottom).

In order to support measurement equivalence, test adaptation goes beyond literal translation to reflect equivalent meaning, format, relevance, intrinsic interest, engagement and familiarity of the item content (Ercikan, 1998[43]; Hambleton, Merenda and Spielberger, 2005[41]). Subtle differences in meaning, difficulty of vocabulary or complexity of sentences between different language versions of tests can affect the difficulty levels of items differentially and lead to incomparability of performances of examinees from different language groups (Allalouf, 2003[44]; Ercikan et al., 2004[45]).

Equivalence of engagement with digital assessments

In addition to potential linguistic differences between different language versions of assessments, the administration of assessments on digital platforms can create sources of non-equivalence in measurement including students' digital literacy (Bennett et al., 2008[46]) and their familiarity with digital platforms (Tate, Warschauer and Abedi, 2016[47]). Digital literacy varies across cultural or social student groups who may have differential levels of access to and experiences with digital devices, which may then affect the extent to which performances may be accounted for by their abilities to use digital devices and navigate through the assessment effectively. Research has demonstrated that digital demands in assessments can affect performance differentially for different cultural or language groups (Ercikan, Asil and Grover, 2018[48]; Fishbein et al., 2018[49]; Zehner et al., 2020[50]).

In particular, requirements for interactivity in assessments heighten the importance of digital literacy in engaging with assessment tasks in ways that support performance. Students from different sociocultural backgrounds may lack familiarity with certain digital platforms or may not use the specific tools and capabilities made available by the assessment similarly (Jackson et al., 2008[51]). In addition, the use of

multimedia formats in some new item types, such as hot-spot items (i.e. identification of correct or incorrect zones) and drag-and-drop image matching, requires that images and videos represent the same meaning and relevance to students from different countries and sociocultural backgrounds (Solano-Flores and Nelson-Barber, 2001[11]). Differential engagement with assessments due to these factors may create sources of incomparability and jeopardise cross-cultural comparability.

Adaptivity

Adaptivity entails tailoring the assessment to students' performance levels in ways that provide measurement efficiency and increase student engagement with the assessment. However, adapting to students' performance levels is also associated with some limitations and challenges (Kingsbury, Freeman and Nesterak, 2014[52]; Yamamoto, Shin and Khorramdel, 2018[53]; Zenisky and Hambleton, 2016[54]). Adaptivity can be at the item level, known as computerised adaptive testing (CAT), or after item sets and at different stages of the assessment, referred to as multi-stage adaptive testing (MST). In both cases, adaptivity in testing relies on the measurement equivalence assumption, that is, that the same constructs are measured by the assessment for different groups and that items have the same order of difficulty for these groups.

Complex constructs such as creativity, communication and collaboration often involve multiple components and measurement dimensions. This multidimensionality not only creates challenges for simple adaptivity designs based on a single dimension but it can also be a source of measurement incomparability. Individuals from different social and cultural backgrounds do not necessarily develop across the construct components in uniform ways. In other words, the dimensionality structure may be different for individuals who may have differential opportunities to develop in different dimensions of the construct. This can result in somewhat different constructs being measured and the ordering of difficulty of items differing for these cultural groups. The assumptions of measurement equivalence, dimensionality and consistency of item ordering are fundamental to establishing cross-cultural comparability in adaptive assessments and these need to be evaluated when designing comparable assessments involving cultural groups. Violations of these expectations can have critical implications, in particular, not meeting the intended goal of measurement efficiency and improving engagement.

There is an additional concern in multi-stage testing (MST) adaptive designs related to the appropriateness of the routing blocks of items and routing decisions for different cultural groups with large degrees of variation in their performance levels. MST adaptivity may start with a block of items determined to have medium-level difficulty across all groups. For groups with much lower performance distribution, a block identified as medium difficulty may be considered difficult or very difficult. This misalignment may also result in MST not meeting its intended goals of advancing measurement efficiency or improving test engagement, in turn affecting the cross-cultural validity and comparability of scores. Therefore in addition to examining measurement equivalence, the appropriateness of an MST design needs to be evaluated for or adapted to cultural groups with varying performance distributions.

Artificial Intelligence-based methodologies in digital assessments

Automated scoring

As educational assessments capture more complex data on student and computer interactions, analyses using machine learning (ML) and AI algorithms have been developed to support automated inferences of student performances (DiCerbo, 2020[55]). AI-based algorithms for scoring, which make use of features generated from natural language processing (NLP) of text, image recognition of visual data or speech recognition of audio data, are critically dependent on the data sources used in the creation of such algorithms (Baker and Hawn, 2021[56]; Manyika, Silberg and Presten, 2019[57]).

In particular, if data sources for these algorithms are restricted to specific cultural and language groups, resulting scores may not have equivalent validity and accuracy for all groups. For example, on an English-speaking test, the AI-based score engine may produce biased scores for students who have different accents and dialects if the model was trained on standard English pronunciations (Benzeghiba et al., 2007[58]). Similar bias may be observed in the automated scoring of text. Previous research has shown that an automated scoring algorithm developed using data from mainstream student groups resulted in less accurate scores for certain racial and ethnic student groups even if responding in the same language (Bridgeman, Trapani and Attali, 2012[59]). This means that automated scoring models need to be trained, calibrated and adjusted using appropriate samples of responses from all target populations (Zhang, 2013[60]), especially when AI-identified features are used in prediction.

As with many AI-based applications, unintended biases can be introduced due to construct-irrelevant features that are correlated with human scores or due to inadequate representation of features that are uniquely observed for minority language or culture groups (Feldman et al., 2015[61]). The ethical use of AI requires that automated scoring systems treat all test takers fairly regardless of language or population groups by providing equivalent meaning of scores for individuals from different social and cultural groups.

An important consideration in the use of automated scoring systems is whether the scoring systems are similarly effective in detecting aberrant responses across languages and cultural groups. Aberrant responses are defined as atypical responses that are not amenable to be scored by algorithms based on most typical response patterns, such as responses that have unusually creative content (e.g. highly metaphorical), exhibit unexpected response organisation (e.g. poem) or have off-topic content (Higgins, Burnstein and Attali, 2006[62]; Zhang, Chen and Ruan, 2015[63]). In multicultural and multilingual assessment contexts, these differences are magnified when students respond in different languages.

Automated item generation

Similar to AI-based automated scoring, automated item generation (AIG) uses a variety of algorithms to automatically create test items. AIG can potentially deliver large amounts of items that cover a variety of content and knowledge, accelerate content updates and test creation, and significantly reduce cost in assessments by replacing highly labour intensive and costly item development by humans. AIG in interactive digital assessments can also be used to generate items "on the fly", in other words, to create and deliver items in interactive and adaptive assessments tailored to students' responses, performance levels and potentially their sociocultural differences.

AIG relies on two processes: an item cognitive model and a computer algorithm that automatically generates items according to the item model. Some risks of using algorithms for generating items have been identified by researchers, such as questionable cognitive models, ill-structured problems that produce multiple correct answers, and implausible or irrelevant distractors (Gierl, Lai and Turner, 2012[64]; Royal et al., 2018[65]). There are additional challenges for multilingual AIG items that are intended to be administered to students from different cultures, where translation quality (awkward phrases, for example) and psychometric properties of the AIG items may not be invariant across cultural groups (Gierl et al., 2016[66]; Higgins, Futagi and Deane, 2005[67]). Hence, two issues require evaluation of the appropriateness of AIG in assessments designed for multicultural and multilingual groups: the first is the degree to which the item cognitive model used for generating items can be assumed to be equivalent for students from different sociocultural contexts; and the other is the linguistic and psychometric equivalence of AIG items generated from different language models trained in different languages.

Test condition equivalence

The final aspect of measurement equivalence is test condition equivalence, which refers to the similarity of test administration conditions such as test instructions, mode and format of the test, timing and advanced

preparations of the test administrators. Large-scale assessments often entail administering the assessments in multiple test sites and across geographical boundaries. The large variability in testing environments and sociocultural norms of learning and assessments therefore increases the threats to score comparability.

Score comparability rests on the premise that valid generalisations can be made about students' performances across test administration sites and cultural and language groups. In multicultural and multilingual assessment contexts, test administrators should be drawn from the local communities so that they are familiar with the culture, language and local dialects to respond to and forestall administration deviations. Training of test administrators is central to the standardisation of testing conditions and should be provided to all administrators in different cultural settings to ensure understanding of the importance of standardised procedures and to provide the needed test administration skills. In the case that the assessments are delivered remotely with live proctors, those proctors need to be trained adequately as well.

When interpreting scores of students from diverse cultural and language backgrounds, knowledge of the broader testing conditions that exist outside assessment settings can enhance understanding of assessment outcomes. These conditions include societal context for testing such as the emphasis given to testing, which may affect how students perceive the testing situation and its role, in turn impacting students' motivation to perform and how they engage with the assessment.

Integrating a sociocultural perspective in assessment design

Despite the recognition that sociocultural context plays an important role in shaping learning and performance on assessment, this context is often neglected in assessment design. In order for assessments to provide equivalent measurement of targeted constructs, the targeted cognition and construct models, task designs and interpretation models (i.e. the three key models of Evidence-Centred Design) need to consider the sociocultural context of learning from the beginning of the design process.

Evaluating construct equivalence

Ensuring cross-cultural validity and comparability needs to start with an evaluation of the equivalence of constructs. This involves identifying behaviours, conceptual understanding and characteristics associated with a construct in a specific culture and time (including how different levels of the competencies involved are differentiated) through surveys or expert judgements (van de Vijver and Tanzer, 2004[68]). The International Test Commission (2017[69]) guidelines emphasise that adequate empirical evidence should be collected to demonstrate that the construct assessed should be understood in the same way across language and cultural groups in large-scale and/or international assessments. In assessments with technological elements, further considerations should be given to how the measurement and scoring of the constructs will be impacted by technology (International Test Commission and Association of Test Publishers, 2022[70]). Gathering such evidence then helps to determine what aspects of the construct can be expected to be common (and what aspects can be expected to differ) across cultural groups considered, and whether a common assessment can provide scores with consistent score meaning across these groups.

Optimising cross-cultural validity

Once the evidence described above has been established, the next step is to develop items for the common and culture-specific elements of the target construct. The common elements will facilitate score comparability and the culture-specific elements will optimise validity of score interpretation and use in different cultural groups. Task models therefore need to consider: 1) whether the definition of the targeted

construct varies for different groups; 2) whether there are expected differences in learning progressions and knowledge structures; 3) whether variations in sociocultural context are expected to lead to different cognitive processes for students from different backgrounds; 4) whether students from different contexts are expected to engage with different features of tasks differently; and 5) what kinds of variations of features of tasks might be needed to optimise performance for students from different sociocultural backgrounds. Multiple versions of tasks may be necessary to obtain equivalent evidence of the targeted construct from different student groups or to redirect the assessment focus to components of the targeted construct where more similarities can be expected for these groups.

Differences in OTL also need to be considered when developing assessment tasks and interpreting and using responses to assessment tasks. In particular, it is important to determine whether students from different population groups can be expected to have had similar opportunities to learn and develop the targeted constructs, what types of tasks might be more closely aligned with their learning experiences, and whether students can be expected to engage with the test environment effectively given their access to similar digital devices, applications and tools. The potential for variations in digital literacy, in particular, to be a source of incomparability can be addressed by intentionally designing assessment tasks for students with the least access to and familiarity with digital resources. When more advanced technology is necessary for an assessment, accessible and effective tutorials can also be provided to help acquaint students with navigating the assessment interface and entering their responses. If possible, practice tasks can also be developed as part of the assessment and distributed to help familiarise students with the digital environment. Ercikan, Asil and Grover (2018[48]) also recommended examining how students from different backgrounds engage with digital assessments using cognitive labs or other response process analyses (see Chapter 12 of this report for a more detailed discussion of the uses of process data for validation purposes).

Adapting assessments into different language versions involves trade-offs between *comparability* and *cultural authenticity*: while concurrent/parallel/simultaneous development (Solano-Flores, Trumbull and Nelson-Barber, 2002[71]) of items in multiple languages prioritises cultural authenticity, successive development (Tanzer and Sim, 1999[72]) of items in one language and then adaptation to other languages prioritises comparability. Having experts evaluate language equivalence is a necessary step as they can identify differences in language, content, format and other aspects of items in the comparison languages. Documenting changes and the rationale for changes between the language versions of assessments is critical for informing test users about the potential impact on comparability.

Validating test scores

Following task design, the next step is interpreting student responses to tasks. There are two components to interpretation models: the *scoring models* used to extract evidence from responses and the *measurement models* used for accumulating evidence across tasks. Scoring models are created based on cognitive theories; however, different world views, knowledge structures and learning progressions might impact student responses. It is therefore important to investigate whether variations in scoring models are needed for obtaining equivalent quality of evidence of the constructs in different student groups.

In general, even when the assessment design integrates all the considerations discussed above, tasks measuring complex skills can remain susceptible to influences from cultural exposure and learning experiences. One possible psychometric solution is the use of universal or anchor items in the test design. Test developers can judiciously identify a set of items that are universally recognised to measure the construct of interest and that can be carefully selected to represent the assessment in terms of assessment content (see Box 11.1 for an example), item statistical characteristics (Kolen and Brennan, 2014[73]) and item formats (Livingston, 2014[74]). These items, to be used as linking items, can be administered to all students. The performance on these universal items can then be used to anchor the differences in performance on the remaining tasks that are more susceptible to influences by cultural norms and

expectations. In this way, universal items can facilitate score comparability and culture-specific item sets can optimise cultural validity of score interpretation and use in different cultural groups.

Further considerations are required for any AI-based automated scoring algorithms, which should be based on construct understanding and evaluated through the lens of validity of interpretation and use (Attali, 2013[75]; Bejar, 2011[76]; Bennett and Bejar, 1998[77]; Powers et al., 2000[78]; Williamson, Xi and Breyer, 2012[79]). This requires systematic evaluation of the consistency of interpretation and use of the AI-based automated scoring engines for individuals from different gender, cultural and language groups, and test developers must ensure that there is an adequate representation of students from all relevant cultural and language groups when training the algorithms. In particular the following set of considerations are important for investigating potential bias in AI-based automated scores (Baker and Hawn, 2021[56]; Benzeghiba et al., 2007[58]; DiCerbo, 2020[55]; ETS, 2021[80]; Kearns et al., 2018[81]):

- Possible human biases in data coding (particularly in supervised learning algorithms).
- Experimenting with different models used in algorithms and paying attention to differential fairness for different cultural groups in the models.
- Cross-validating the models with new and different datasets.
- Investigating potential bias by comparing human versus machine scores for different cultural groups.

In addition, it is necessary to evaluate the equivalency of NLP features that feed into AI algorithms across languages and investigate whether the same, different or hierarchical AI-based scoring models should be used for different cultural and language groups (see McCaffrey et al. (2022[82]) for a more detailed discussion of these issues). The involvement of subject experts from different cultural and language groups in developing AI-based scoring (and automated item generation) is necessary for minimising poor quality item development and incomparability across language and cultural groups.

Methodologies for examining equivalence

While several steps should be taken throughout the assessment design process to establish evidence to support construct and test equivalence, these issues also need to be evaluated by large-scale psychometric studies using various methodologies. The most commonly used methodology for examining measurement invariance at the scale or test level is confirmatory factor analysis (CFA) that compares test data structures across comparison groups (Ercikan and Koh, 2005[40]; Oliveri and Ercikan, 2011[83]). At the item level, differential item functioning (DIF) analysis evaluates whether the probability of a correct response among equally able students is the same for comparison groups (Guo and Dorans, 2019[84]; 2020[85]; Dorans and Holland, 1993[86]; Holland and Thayer, 1988[87]) and it has been the psychometric approach for examining measurement equivalence across groups since the 1980s (Dorans and Holland, 1993[86]; Holland and Thayer, 1988[87]; Holland and Wainer, 1993[88]; Rogers and Swaminathan, 2016[89]; Shealy and Stout, 1993[90]). Research indicates that measurement incomparability identified at the item level does not necessarily result in observable test or scale level differences (Ercikan and Gonzalez, 2008[91]; Zumbo, 2003[92]). This highlights the importance of examining factor structure equivalence at both the test and item levels to provide complimentary evidence for a full evaluation of measurement invariance.

In large-scale assessments that involve large numbers of language and cultural groups, various other methods may be used to examine measurement invariance across groups. In particular in international assessments, measurement invariance is determined by the item-by-language (country) interaction in the item parameter estimates (see Chapters 9 and 16 in OECD (2017[93]), for example). Group-specific parameters (i.e. country item parameters) for items exhibiting group-level DIF in the international calibration are estimated to reduce potential bias introduced by these interactions. If multiple language-adapted assessments are produced, then a linking study may also be needed to create comparable scales

with measurement unit equivalence. Research on comparability (Sireci, 1997[94]; 2005[95]) indicates that, in the absence of sufficient evidence for measurement equivalence across groups, score scales should be based on separate language/country calibrations and comparability should be established through a linking procedure.

Interpretation of DIF findings

Several considerations must be kept in mind for interpreting statistical findings in the context of measurement invariance research. First, some level of incomparability exists when measurement is compared for all assessment groups and statistical significance in the violation of measurement invariance may not be a useful indicator in evaluating its practical consequences, especially when sample sizes are large. Effect size measures provide a better indication of the level of incomparability (Nye and Drasgow, 2011[96]) to facilitate making decisions about the exclusion of items or the revision of scales to establish comparability.

Second, statistical results do not always guide what actions should be taken if there is evidence of measurement variance and incomparability. Recent studies examining and confirming sources of DIF have advocated for the use of mixed methods approaches that integrate quantitative results from DIF analysis and qualitative findings from expert appraisal to uncover sources of DIF across comparison groups (Benítez and Padilla, 2014[97]; Benítez et al., 2016[98]). Ercikan et al. (2010[99]) demonstrated that Think Aloud Protocols (TAPs) could be used as an approach for examining and confirming sources of DIF in multiple language versions of assessments. Digital assessment environments also provide opportunities for examining measurement equivalence in different ways by providing information about student behaviour and cognitive processes in the data logs that can be used for examining the comparability of response processes and patterns for students from different cultural groups (Ercikan and Pellegrino, 2017[100]; Guo and Ercikan, 2021[19]) – see also Chapter 12 of this report for more detail.

Third, most current DIF methodologies are designed for the comparison of pre-specified focal and reference groups (i.e. observable manifest groups characterised, for example, by gender or ethnicity). In other words, DIF methods do not identify hidden bias for latent heterogeneity groups that might nonetheless exist. For instance, the assumption of within-group homogeneity often neglects the actual heterogeneity that exists in subgroups (Cohen and Bolt, 2005[101]; Ercikan and Por, 2020[36]; Grover and Ercikan, 2017[102]). Oliveri, Ercikan and Zumbo (2014[103]) demonstrated in a simulation study that an increase in heterogeneity from 0 to 80 percent within the focal groups decreased the accuracy of DIF detection.

Different approaches have been used to account for the heterogeneity in focal and reference groups used in DIF analyses. One such approach, referred to as “melting pot” DIF (Dorans and Holland, 1993[86]) or DIF dissection approach (Zhang, Dorans and Matthews-López, 2005[104]), focused on crossing two manifest groups (e.g. gender and ethnicity) to create more specific subgroups for analysis. Other approaches focused on identifying latent homogenous groups through statistical analyses (Cohen and Bolt, 2005[101]; Strobl, Kopf and Zeileis, 2015[105]). Ercikan and Oliveri (2013[106]) also proposed a two-step approach in conducting DIF using latent class analysis within manifest groups: the first step involves a latent class analysis to identify heterogeneous groupings in the considered populations, and the second step involves applying DIF methodologies to the identified latent classes rather than manifest groups as a whole.

Conclusion

Assessing complex constructs using engaging tasks, often on digital-based platforms, is critical for promoting learning and development of high value skills, knowledge and competencies, and necessary for advancing assessment methodologies. In this chapter we argued for the recognition of the complex

sociocultural context that assessments are conducted in and the importance of cross-cultural validity and comparability. In particular, assessment designers need to take the complex sociocultural context into account in deciding what to assess, how to assess it and how assessment results need to be interpreted and used. We highlighted key measurement equivalence issues that arise specifically in digital assessments of complex constructs in multicultural populations. Many of these issues can be mitigated through a principled assessment design process that examines sociocultural influences at the onset of defining the assessment constructs, designing tasks, and developing scoring and measurement models. However, even when all these are taken into account, empirical investigations and supporting empirical evidence are necessary for establishing the validity and comparability of assessment results for individuals from different cultural and language groups.

References

AERA, APA, NCME (2014), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, D.C., https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf. [31]

Allalouf, A. (2003), "Revising translated differential item functioning items as a tool for improving cross-lingual assessment", *Applied Measurement in Education*, Vol. 16/1, pp. 55-73, https://doi.org/10.1207/s15324818ame1601_3. [44]

Allalouf, A., R. Hambleton and S. Sireci (1999), "Identifying the causes of DIF in translated verbal items", *Journal of Educational Measurement*, Vol. 36/3, pp. 185-198, https://doi.org/10.1111/j.1745-3984.1999.tb00553.x. [38]

Attali, Y. (2013), "Validity and reliability of automated essay scoring", in Shermis, M. and J. Burstein (eds.), *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, Routledge, New York, https://doi.org/10.4324/9780203122761. [75]

Baker, R. and A. Hawn (2021), "Algorithmic bias in education", *International Journal of Artificial Intelligence in Education*, Vol. 32/4, pp. 1052-1092, https://doi.org/10.1007/s40593-021-00285-9. [56]

Bejar, I. (2011), "A validity-based approach to quality control and assurance of automated scoring", *Assessment in Education: Principles, Policy & Practice*, Vol. 18/3, pp. 319-341, https://doi.org/10.1080/0969594x.2011.555329. [76]

Benítez, I. and J. Padilla (2014), "Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing", *Journal of Mixed Methods Research*, Vol. 8/1, pp. 52-68, https://doi.org/10.1177/1558689813488245. [97]

Benítez, I. et al. (2016), "Using mixed methods to interpret differential item functioning", *Applied Measurement in Education*, Vol. 29/1, pp. 1-16, https://doi.org/10.1080/08957347.2015.1102915. [98]

Bennett, R. and I. Bejar (1998), "Validity and automad scoring: It's not only the scoring", *Educational Measurement: Issues and Practice*, Vol. 17/4, pp. 9-17, https://doi.org/10.1111/j.1745-3992.1998.tb00631.x. [77]

Bennett, R. et al. (2008), "Does it matter if I take my mathematics test on computer? A second empirical study of mode effects in NAEP", *Journal of Technology, Learning, and Assessment*, http://www.jtla.org (accessed on 4 March 2023). [46]

Benzeghiba, M. et al. (2007), "Automatic speech recognition and speech variability: A review", *Speech Communication*, Vol. 49/10-11, pp. 763-786, https://doi.org/10.1016/j.specom.2007.02.006. [58]

Bergqvist, E., F. Theens and M. Österholm (2018), "The role of linguistic features when reading and solving mathematics tasks in different languages", *The Journal of Mathematical Behavior*, Vol. 51, pp. 41-55, https://doi.org/10.1016/j.jmathb.2018.06.009. [42]

Bridgeman, B., C. Trapani and Y. Attali (2012), "Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country", *Applied Measurement in Education*, Vol. 25/1, pp. 27-40, https://doi.org/10.1080/08957347.2012.635502. [59]

Cogan, L., W. Schmidt and D. Wiley (2001), "Who takes what math and in which track? Using TIMSS to characterize U.S. students' eighth-grade mathematics learning opportunities", *Educational Evaluation and Policy Analysis*, Vol. 23/4, pp. 323-341, https://doi.org/10.3102/01623737023004323. [26]

Cohen, A. and D. Bolt (2005), "A mixture model analysis of differential item functioning", *Journal of Educational Measurement*, Vol. 42/2, pp. 133-148, https://doi.org/10.1111/j.1745-3984.2005.00007. [101]

Cooper, C. et al. (2009), "Poverty, race, and parental involvement during the transition to elementary school", *Journal of Family Issues*, Vol. 31/7, pp. 859-883, https://doi.org/10.1177/0192513x09351515. [14]

DiCerbo, K. (2020), "Assessment for learning with diverse learners in a digital world", *Educational Measurement: Issues and Practice*, Vol. 39/3, pp. 90-93, https://doi.org/10.1111/emip.12374. [55]

Dorans, N. and P. Holland (1993), "DIF detection and description: Mantel-Haenszel and standardization", in Holland, P. and H. Wainer (eds.), *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale. [86]

El Masri, Y., J. Baird and A. Graesser (2016), "Language effects in international testing: The case of PISA 2006 science items", *Assessment in Education: Principles, Policy & Practice*, Vol. 23/4, pp. 427-455, https://doi.org/10.1080/0969594x.2016.1218323. [39]

Ercikan, K. (1998), "Translation effects in international assessments", *International Journal of Educational Research*, Vol. 29/6, pp. 543-553, https://doi.org/10.1016/s0883-0355(98)00047-0. [43]

Ercikan, K. et al. (2010), "Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews", *Educational Measurement: Issues and Practice*, Vol. 29/2, pp. 24-35, https://doi.org/10.1111/j.1745-3992.2010.00173.x. [99]

Ercikan, K., M. Asil and R. Grover (2018), "Digital divide: A critical context for digitally based assessments", *Education Policy Analysis Archives*, Vol. 26/51, pp. 1-24, https://doi.org/10.14507/epaa.26.3817. [48]

Ercikan, K. and S. Elliott (2015), "Assessment as a tool for communication and improving educational equity", *A white paper for the Smarter Balanced Assessment Consortium*. [13]

Ercikan, K. et al. (2004), "Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests", *Applied Measurement in Education*, Vol. 17/3, pp. 301-321, https://doi.org/10.1207/s15324818ame1703_4. [45]

Ercikan, K. and E. Gonzalez (2008), "Score scale comparability in international assessments", *Paper presented at the National Council on Measurement in Education, New York*. [91]

Ercikan, K., H. Guo and Q. He (2020), "Use of response process data to inform group comparisons and fairness research", *Educational Assessment*, Vol. 25/3, pp. 179-197, https://doi.org/10.1080/10627197.2020.1804353. [18]

Ercikan, K. and K. Koh (2005), "Examining the construct comparability of the English and French versions of TIMSS", *International Journal of Testing*, Vol. 5/1, pp. 23-35, https://doi.org/10.1207/s15327574ijt0501_3. [40]

Ercikan, K. and J. Lyons-Thomas (2013), "Adapting tests for use in other languages and cultures", in Geisinger, K. et al. (eds.), *APA Handbook of Testing and Assessment in Psychology, Vol. 3: Testing and Assessment in School Psychology and Education*, American Psychological Association, Washington, D.C., https://doi.org/10.1037/14049-026. [6]

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. [7]

Ercikan, K. and M. Oliveri (2013), "Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations", in Chatterji, M. (ed.), *Validity, Fairness and Testing of Individuals in High Stakes Decision-Making Context*, Emerald Publishing, Bingley. [106]

Ercikan, K. and J. Pellegrino (2017), "Validation of score meaning using examinee response processes for the next generation of assessments", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, https://doi.org/10.4324/9781315708591. [100]

Ercikan, K. and H. Por (2020), "Comparability in multilingual and multicultural assessment contexts", in Berman, A., E. Haertel and J. Pellegrino (eds.), *Comparability in Large-Scale Assessment: Issues and Recommendations*, National Academy of Education, Washington, D.C., https://naeducation.org/wp-content/uploads/2020/04/8-Comparability-in-Multilingual-and-Multicultural-Assessment-Contexts.pdf. [36]

Ercikan, K., W. Roth and M. Asil (2015), "Cautions about inferences from international assessments: The case of PISA 2009", *Teachers College Record*, Vol. 117/1, pp. 1-28, https://doi.org/10.1177/016146811511700107. [20]

Ercikan, K. et al. (2014), "Inconsistencies in DIF detection for sub-groups in heterogeneous language groups", *Applied Measurement in Education*, Vol. 27/4, pp. 273-285, https://doi.org/10.1080/08957347.2014.944306. [12]

ETS (2021), *Best Practices for Constructed-Response Scoring*, Educational Testing Service, https://www.ets.org/pdfs/about/cr_best_practices.pdf (accessed on 4 March 2023). [80]

Feldman, M. et al. (2015), "Certifying and removing disparate impact", *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259-268, https://doi.org/10.1145/2783258.2783311. [61]

Fishbein, B. et al. (2018), "The TIMSS 2019 Item Equivalence Study: Examining mode effects for computer-based assessment and implications for measuring trends", *Large-scale Assessments in Education*, Vol. 6/1, https://doi.org/10.1186/s40536-018-0064-z. [49]

Foster, N. and A. Schleicher (2021), "Assessing creative skills", *Creative Education*, Vol. 13, pp. 1-29, https://doi.org/10.4236/ce.2022.131001. [33]

Fuchs, T. and L. Wößmann (2007), "What accounts for international differences in student performance? A re-examination using PISA data", *Empirical Economics*, Vol. 32, pp. 433-464, https://doi.org/10.1007/s00181-006-0087-0. [27]

Gee, J. (2008), "A sociocultural perspective on opportunity to learn", in Moss, P. et al. (eds.), *Assessment, Equity, and Opportunity to Learn*, Cambridge University Press, Cambridge, https://doi.org/10.1017/CBO9780511802157.004. [2]

Gierl, M. et al. (2016), "Using technology-enhanced processes to generate test items in multiple languages", in Drasgow, F. (ed.), *Technology and Testing: Improving Educational and Psychological Measurement*, Routledge, New York, https://doi.org/10.4324/9781315871493. [66]

Gierl, M., H. Lai and S. Turner (2012), "Using automatic item generation to create multiple-choice test items", *Medical Education*, Vol. 46/8, pp. 757-765, https://doi.org/10.1111/j.1365-2923.2012.04289.x. [64]

Grover, R. and K. Ercikan (2017), "For which boys and which girls are reading assessment items biased against? Detection of differential item functioning in heterogeneous gender populations", *Applied Measurement in Education*, Vol. 30/3, pp. 178-195, https://doi.org/10.1080/08957347.2017.1316276. [102]

Guo, H. and N. Dorans (2020), "Using weighted sum scores to close the gap between DIF practice and theory", *Journal of Educational Measurement*, Vol. 57/4, pp. 484-510, https://doi.org/10.1111/jedm.12258. [85]

Guo, H. and N. Dorans (2019), "Observed scores as matching variables in differential item functioning under the one- and two-parameter logistic models: Population results", *ETS Research Report Series*, Vol. 2019/1, pp. 1-27, https://doi.org/10.1002/ets2.12243. [84]

Guo, H. and K. Ercikan (2021), "Differential rapid responding across language and cultural groups", *Educational Research and Evaluation*, Vol. 26/5-6, pp. 302-327, https://doi.org/10.1080/13803611.2021.1963941. [19]

Hambleton, R., P. Merenda and C. Spielberger (eds.) (2005), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Psychology Press, New York, https://doi.org/10.4324/9781410611758. [41]

Higgins, D., J. Burnstein and Y. Attali (2006), "Identifying off-topic student essays without topic-specific training data", *Natural Language Engineering*, Vol. 12/2, pp. 145-159, https://doi.org/10.1017/s1351324906004189. [62]

Higgins, D., Y. Futagi and P. Deane (2005), "Multilingual generalization of the ModelCreator software for math item generation", *ETS Research Report Series*, Vol. 2005/1, pp. i-38, https://doi.org/10.1002/j.2333-8504.2005.tb01979.x. [67]

Holland, P. and D. Thayer (1988), "Differential item performance and the Mantel-Haenszel procedure", in Wainer, H. and H. Braun (eds.), *Test Validity*, Lawrence Erlbaum, Hillsdale. [87]

Holland, P. and H. Wainer (1993), *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale. [88]

International Test Commission (2017), "ITC guidelines for translating and adapting tests (Second edition)", *International Journal of Testing*, Vol. 18/2, pp. 101-134, https://doi.org/10.1080/15305058.2017.1398166. [69]

International Test Commission and Association of Test Publishers (2022), *Guidelines for Technology-Based Assessment*, https://www.intestcom.org/page/16 (accessed on 4 March 2023). [70]

Jackson, L. et al. (2008), "Race, gender, and information technology use: The new digital divide", *CyberPsychology & Behavior*, Vol. 11/4, pp. 437-442, https://doi.org/10.1089/cpb.2007.0157. [51]

Kearns, M. et al. (2018), "Preventing fairness gerrymandering: Auditing and learning for subgroups fairness", in Dy, J. and A. Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, PMLR 80*, http://proceedings.mlr.press/v80/kearns18a/kearns18a.pdf. [81]

Kingsbury, G., E. Freeman and M. Nesterak (2014), "The potential of adaptive assessment", *Educational Leadership*, Vol. 71/6, pp. 12-18. [52]

Kolen, M. and R. Brennan (2014), *Test Equating, Scaling, and Linking: Methods and Practices*, Springer, New York, https://doi.org/10.1007/978-1-4939-0317-7. [73]

Kurlaender, M. and J. Yun (2007), "Measuring school racial composition and student outcomes in a multiracial society", *American Journal of Education*, Vol. 113/2, pp. 213-243, https://doi.org/10.1086/510166. [30]

Kurlaender, M. and J. Yun (2005), "Fifty years after Brown: New evidence of the impact of school racial composition on student outcomes", *International Journal of Educational Policy, Research, and Practice: Reconceptualizing Childhood Studies*, Vol. 6/1, pp. 51-78. [29]

Lee, J. and L. Stankov (2018), "Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA", *Learning and Individual Differences*, Vol. 65, pp. 50-64, https://doi.org/10.1016/j.lindif.2018.05.009. [15]

Lipka, J. and T. McCarty (1994), "Changing the culture of schooling: Navajo and Yup'ik cases", *Anthropology & Education Quarterly*, Vol. 25/3, pp. 266-284, https://doi.org/10.1525/aeq.1994.25.3.04x0144n. [4]

Liu, Y., A. Wu and B. Zumbo (2006), "The relation between outside of school factors and mathematics achievement: A cross-country study among the US and five top-performing Asian countries", *Journal of Educational Research & Policy Studies*, Vol. 6, pp. 1-35. [10]

Livingston, S. (2014), *Equating Test Scores (Without IRT)*, https://www.ets.org/Media/Research/pdf/LIVINGSTON2ed.pdf (accessed on 4 March 2023). [74]

Lubart, T. (1990), "Creativity and cross-cultural variation", *International Journal of Psychology*, Vol. 25/1, pp. 39-59, https://doi.org/10.1080/00207599008246813. [9]

Manyika, J., J. Silberg and B. Presten (2019), *What do we do about the biases in AI?*, https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai (accessed on 4 March 2023). [57]

McCaffrey, D. (2022), "Best practices for constructed-response scoring", *ETS Research Report Series*, Vol. 2022/1, pp. 1-58, https://doi.org/10.1002/ets2.12358. [82]

Moss, P. et al. (eds.) (2008), *Assessment, Equity, and Opportunity to Learn*, Cambridge University Press, Cambridge, https://doi.org/10.1017/cbo9780511802157. [3]

National Academies of Sciences, Engineering, and Medicine (2019), *Monitoring Educational Equity*, National Academies Press, Washington, D.C., https://doi.org/10.17226/25389. [21]

National Center for Educational Statistics (2021), *Survey questionnaires: Questionnaires for students, teachers, and school administrators*, https://nces.ed.gov/nationsreportcard/experience/survey_questionnaires.aspx (accessed on 4 March 2023). [22]

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. [8]

Nye, C. and F. Drasgow (2011), "Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups", *Journal of Applied Psychology*, Vol. 96/5, pp. 966-980, https://doi.org/10.1037/a0022955. [96]

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). [34]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris, https://www.oecd.org/pisa/data/2015-technical-report/ (accessed on 4 March 2023). [93]

OECD (2016), *Equations and Inequalities: Making Mathematics Accessible to All*, OECD Publishing, Paris, https://doi.org/10.1787/9789264258495-en. [23]

Oliveri, M. and K. Ercikan (2011), "Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions?", *Applied Measurement in Education*, Vol. 24/4, pp. 349-366, https://doi.org/10.1080/08957347.2011.607063. [83]

Oliveri, M., K. Ercikan and B. Zumbo (2014), "Effects of population heterogeneity on accuracy of DIF detection", *Applied Measurement in Education*, Vol. 27/4, pp. 286-300, https://doi.org/10.1080/08957347.2014.944305. [103]

Pellegrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know*, National Academies Press, Washington, D.C., https://doi.org/10.17226/10019. [1]

Perie, M. (2020), "Comparability across different assessment systems", in Berman, A., E. Haertel and J. Pellegrino (eds.), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, National Academy of Education, Washington, D.C., https://doi.org/10.31094/2020/1. [32]

Powers, D. et al. (2000), "Comparing the validity of automated and human essay scoring", *ETS Research Report Series*, Vol. 2000/2, pp. i-23, https://doi.org/10.1002/j.2333-8504.2000.tb01833.x. [78]

Rogers, H. and H. Swaminathan (2016), "Concepts and methods in research on differential functioning of test items: Past, present, and future", in Wells, C. and M. Faulkner-Bond (eds.), *Educational Measurement: From Foundations to Future*, The Guilford Press, New York. [89]

Rotberg, I. (2006), "Assessment around the world", *Educational Leadership*, Vol. 64/3, pp. 58-63, https://neqmap.bangkok.unesco.org/wp-content/uploads/2019/08/Assessment-Around-the-World.pdf. [16]

Royal, K. et al. (2018), "Automated item generation: The future of medical education assessment", *EMJ Innovations*, Vol. 2/1, pp. 88-93, https://doi.org/10.33590/emjinnov/10313113. [65]

Schleicher, A. (2019), *PISA 2018: Insights and Interpretations*, https://www.oecd.org/pisa/PISA%202018%20Insights%20and%20Interpretations%20FINAL%20PDF.pdf (accessed on 4 March 2023). [24]

Schmidt, W. et al. (2015), "The role of schooling in perpetuating educational inequality", *Educational Researcher*, Vol. 44/7, pp. 371-386, https://doi.org/10.3102/0013189x15603982. [28]

Schmidt, W., P. Zoido and L. Cogan (2014), "Schooling matters: Opportunity to learn in PISA 2012", *OECD Education Working Papers* No. 95, https://doi.org/10.1787/19939019. [25]

Shealy, R. and W. Stout (1993), "A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF", *Psychometrika*, Vol. 58/2, pp. 159-194, https://doi.org/10.1007/BF02294572. [90]

Sireci, S. (2005), "Using bilinguals to evaluate the comparability of different language versions of a test", in Hambleton, R., P. Merenda and C. Spielberger (eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Lawrence Erlbaum, Hillsdale. [95]

Sireci, S. (1997), "Problems and issues in linking assessments across languages", *Educational Measurement: Issues and Practice*, Vol. 16/1, pp. 12-19, https://doi.org/10.1111/j.1745-3992.1997.tb00581.x. [94]

Solano-Flores, G. and S. Nelson-Barber (2001), "On the cultural validity of science assessments", *Journal of Research in Science Teaching*, Vol. 38/5, pp. 553-573, https://doi.org/10.1002/tea.1018. [11]

Solano-Flores, G., E. Trumbull and S. Nelson-Barber (2002), "Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities", *International Journal of Testing*, Vol. 2/2, pp. 107-129, https://doi.org/10.1207/s15327574ijt0202_2. [71]

Sternberg, R. (2013), "Intelligence", in Freedheim, D. and I. Weiner (eds.), *Handbook of Psychology: History of Psychology*, John Wiley & Sons, Hoboken. [35]

Strobl, C., J. Kopf and A. Zeileis (2015), "Rasch trees: A new method for detecting differential item functioning in the Rasch model", *Psychometrika*, Vol. 80/2, pp. 289-316, https://doi.org/10.1007/s11336-013-9388-3. [105]

Suzuki, L. and J. Ponterotto (2007), *Handbook of Multicultural Assessment: Clinical, Psychological, and Educational Applications*, John Wiley & Sons, Hoboken. [5]

Tanzer, N. and C. Sim (1999), "Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations", *European Journal of Psychological Assessment*, Vol. 15/3, pp. 258-269, https://doi.org/10.1027//1015-5759.15.3.258. [72]

Tate, T., M. Warschauer and J. Abedi (2016), "The effects of prior computer use on computer-based writing: The 2011 NAEP writing assessment", *Computers & Education*, Vol. 101, pp. 115-131, https://doi.org/10.1016/j.compedu.2016.06.001. [47]

van de Vijver, F. and N. Tanzer (2004), "Bias and equivalence in cross-cultural assessment: an overview", *European Review of Applied Psychology*, Vol. 54/2, pp. 119-135, https://doi.org/10.1016/j.erap.2003.12.004. [68]

van de Vijver, F. and N. Tanzer (1997), "Bias and equivalence in cross-cultural assessment: An overview", *European Review of Applied Psychology*, Vol. 47/4, pp. 263-280. [37]

Williamson, D., X. Xi and F. Breyer (2012), "A framework for evaluation and use of automated scoring", *Educational Measurement: Issues and Practice*, Vol. 31/1, pp. 2-13, https://doi.org/10.1111/j.1745-3992.2011.00223.x. [79]

Yamamoto, K., H. Shin and L. Khorramdel (2018), "Multistage adaptive testing design in international large-scale assessments", *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 16-27, https://doi.org/10.1111/emip.12226. [53]

Zehner, F. et al. (2020), "PISA reading: Mode effects unveiled in short text responses", *Psychological Test and Assessment Modeling*, Vol. 62/1, pp. 85-105, https://doi.org/10.25656/01:20354. [50]

Zenisky, A. and R. Hambleton (2016), "Multi-stage test design: Moving research results into practice", in Yan, D., A. von Davier and C. Lewis (eds.), *Computerized Multistage Testing*, Chapman and Hall/CRC, New York. [54]

Zhang, H. and F. Luo (2020), "The development of psychological and educational measurement in China", *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*, https://www.ce-jeme.org/journal/vol1/iss1/7 (accessed on 4 March 2023). [17]

Zhang, M. (2013), "The impact of sampling approach on population invariance in automated scoring of essays", *ETS Research Report Series*, Vol. 2013/1, pp. i-33, https://doi.org/10.1002/j.2333-8504.2013.tb02325.x. [60]

Zhang, M., J. Chen and C. Ruan (2015), "Evaluating the detection of aberrant responses in automated essay scoring", in van der Ark, L. et al. (eds.), *Quantitative Psychology Research. Springer Proceedings in Mathematics & Statistics*, Springer, Cham, https://doi.org/10.1007/978-3-319-19977-1_14. [63]

Zhang, Y., N. Dorans and J. Matthews-López (2005), "Using DIF dissection method to assess effects of item deletion", *ETS Research Report Series*, Vol. 2005/2, pp. i-11, https://doi.org/10.1002/j.2333-8504.2005.tb02000.x. [104]

Zumbo, B. (2003), "Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests", *Language Testing*, Vol. 20/2, pp. 136-147, https://doi.org/10.1191/0265532203lt248oa. [92]

12 Uses of process data in advancing the practice and science of technology-rich assessments

By Kadriye Ercikan, Hongwen Guo and Han Hui Por

(Educational Testing Service)

Digital assessments and technological advances provide opportunities for significant innovations in assessments, including assessing complex constructs where both the solution and the response processes are important assessment targets. This chapter discusses how technology-rich assessments can capture data representing test takers' response processes at a large scale that may provide insights on students' knowledge and abilities and how they interact and engage with assessments. It discusses three key uses of process data captured on digital platforms during the test-taking processes, including for improving assessment design and score quality and validity, for providing evidence of the targeted construct, and for investigating group comparisons and fairness. The chapter describes each of these key uses and appropriate methodologies for optimising the use of process data in digital assessments.

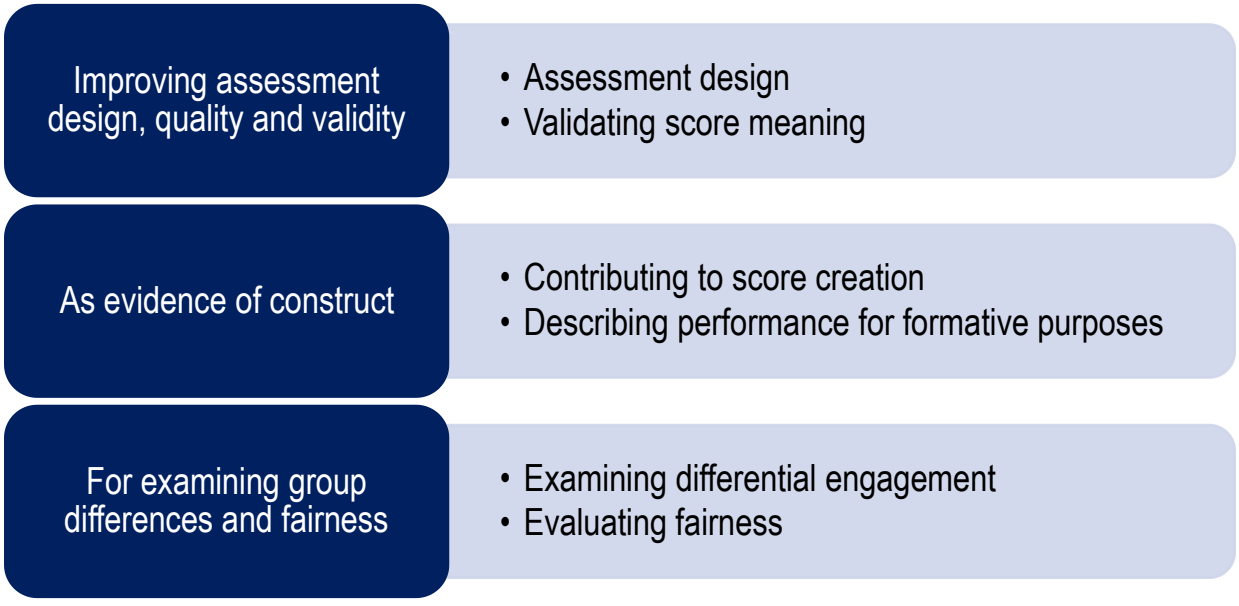
Introduction

Increasingly, assessment design and development are informed by and provide information about student learning processes. This is motivated by the need to understand how students learn, to align assessments with cognition models, and to provide feedback to learners and teachers. In addition, there is a growing emphasis on the importance of assessing complex constructs such as problem solving and creative thinking, which require capturing response processes in addition to the final solutions to assessment tasks. Technology-rich assessments provide opportunities for presenting items in engaging contexts that may include audio, videos, images, text, interactivity, and include other dynamic displays and digital tools. These opportunities are accompanied by advancements in data capture capabilities that allow for collecting data on response processes including eye-tracking, click streams and time stamps associated with different actions and events (also refer to Table 7.1 and Table 10.1 in Chapters 7 and 10 of this volume, respectively).

Research has demonstrated the benefits of multimedia item formats in increasing engagement as well as their potential negative impact on validity due to a multitude of cognitive demands not targeted by the assessment (McNamara et al., 2011[1]; Popp, Tuzinski and Fetzer, 2015[2]); see also Chapter 11 of this volume. Despite the potential for contributing to incomparability, digital assessment environments provide opportunities for capturing response process data that can advance measurement in significant ways from assessment design, item development, validation of score meaning and use, and for group comparisons and fairness research. Use of response process data requires interpretation of clicks, timing and other action data with respect to what they may indicate in cognitive thinking processes.

In the first section of this chapter, we discuss approaches to creating response process indicators using process data, focusing on timing data and using rapid guessing behaviour indicators as an example. This is followed by three sections highlighting three key, interrelated uses of process data (Figure 12.1): 1) to improve assessment design, quality and validity; 2) as evidence of the targeted construct; and 3) for examining differential engagement in the assessment by different groups, for score comparability and detecting construct-irrelevant variables (CIVs). The chapter ends by highlighting the importance of using process data during the initial design stages of technology-rich assessments.

Figure 12.1. Three key uses of process data



Construction of response process indicators

One of the key steps in utilising process data is identifying aspects of data that might be meaningful in reflecting cognitive response processes – that is, creating response process indicators. Descriptions of such indicators, inspired by Intelligent Tutoring Systems (ITS), can be found in Chapter 10 of this volume as well as reviews by Baker et al. (2019[3]) and Sottilare et al. (2013[4]). This section focuses on process indicators extracted from process data captured in educational assessments. We discuss methodologies for creating response process indicators and their uses in analysing whether students are engaging with tasks in intended ways as a valid measure of their ability or exhibiting unintended behaviours such as rapid guessing.

There are two approaches to creating such indicators: *top-down* and *bottom-up*. The top-down approach for creating process indicators is theory-guided and uses a known cognitive theory (i.e. one that experts agree upon in the literature and practice) to determine what evidence in the log data reflects students' cognitive thinking processes and how to extract such evidence. Kroehne and Goldhammer (2018[5]) proposed a generic framework – finite state machines – to analyse log data using a top-down approach. States represent parts of a theoretically defined response process consisting of filtered and/or integrated log data. Theoretical considerations constitute the meaning of states – for example, students are expected to read the question stem and response options before providing the first response to an item set. In this example, the sequence of states decomposes the problem-solving process to provide the theoretical foundation for defining process indicators. Kroehne and Goldhammer (2018[5]) also illustrated how item-level time components, states and indicators (such as time on reading item stem, time on solving question) were extracted from log data for individual questions in an item set presented on one screen and how they could be used to explore the relationship between response times and item responses. In another study, Goldhammer et al. (2021[6]) created the item-level indicator, which reflects the construct allocation of cognitive resources and is presented by the time spent on relevant pages.

Guided by cognitive writing theories, Guo and colleagues (2019[7]; 2020[8]) also used a top-down approach to extract information from log data. They defined *a priori* finite states (e.g. long pause, editing, text production and global editing) extracted from keystroke data, with time stamps, in essay writing processes. Markov modelling of the state sequences and their duration times showed group differences in the writing processes when given comparable essay score distributions. Writing states can also be aggregated into process measures, such as frequencies in different writing stages, to investigate students' writing styles (Zhang, Guo and Liu, 2021[9]). Bennett and colleagues (2021[10]) further used writing process features to create higher-level process indicators to investigate different writing styles among students with different writing proficiency.

In contrast, bottom-up or data-driven approaches involve the creation of indicators of behaviour types using exploratory data analysis methods, such as cluster analysis, data visualisation and unsupervised or semi-supervised learning. Researchers identify patterns or clusters in the data and then validate and interpret the findings against cognitive theories or external evidence (Fratamico et al., 2017[11]; Perez et al., 2017[12]). Process indicators can also be created using a combination of top-down and bottom-up approaches. Guided by cognitive theory and the data exploration literature, Greiff and colleagues (2016[13]) examined log files collected from complex problem-solving tasks. They studied three process indicators: time on task, non-interfering observation (time before the first action), and intervention frequency (interaction events with the test environment), and their relationships with performance. Their analysis of the behavioural data helped to understand students' performance.

In the following sub-section, we use the rapid guessing behaviour indicator as a concrete example to explain the creation of a process indicator. As discussed in Chapter 9 in this volume, triangulation between top-down and bottom-up approaches can provide evidence supporting the interpretation of different response processes. A similar approach that strikes a balance between guidance by theory and exploration

of data was discussed by Fratamic and colleagues (2017[11]) in an adaptive learning context using interactive simulations.

It is important to highlight that indicators developed through both top-down and bottom-up approaches require validation of their intended interpretation and use, meaning theoretical and/or empirical rationales are needed to support the validity of inferences about (latent) attributes of test takers' response processes (Fratamico et al., 2017[11]; Goldhammer et al., 2021[6]). Validity of interpretation of process indicators was examined by Goldhammer and colleagues (2021[6]) using correlational and experimental validation strategies. The authors explained the process of reasoning from log data to low-level features and then process indicators as the outcome of evidence identification, where contextualising information from log data is essential. Such exploration and validation studies help build process models that can guide test development. Although it is not generally possible to develop formal process models and indicators that will account for the full range of performances and behaviours in the domain, it is often possible to develop such models for specific kinds of test tasks in the target domain in a technology-rich environment (Kane and Mislevy, 2017[14]). Validation of process indicators plays a critical role in evaluating their meaningful application in measurement.

Process indicator for rapid guessing behaviour

Valid interpretation and uses of assessment results require item responses to be indicators of the targeted constructs. When students engage with items in ways that do not reflect their abilities and competencies, such behaviours may compromise validity. A primary example of such behaviours is rapid guessing. Previous research has commonly used item response time (Buchholz, Cignetti and Piacentini, 2022[15]; Wise, 2017[16]; Wise, 2021[17]) to identify rapid responding behaviours. Rapid guessing behaviour is usually identified as students' lack of meaningful engagement with an assessment situation such that they respond to assessment items rapidly without taking the time to read and fully consider the items. It is also hypothesised that, for a multiple-choice item, rapid guessing responses on average lead to a chance score, which may not reflect students' true knowledge and skills (DeMars and Wise, 2010[18]; Kroehne, Deribo and Goldhammer, 2020[19]; Schnipke and Scrams, 2002[20]; Wise, 2017[16]). In intelligent learning or tutoring systems, such effortless behaviours are relevant to exploiting hints and “gaming the system” and are associated with lower learning gains (see, for example, Baker et al. (2019[3]) and Roll et al. (2005[21])).

Many procedures have been developed to identify rapid guessing responses (Wise, 2017[16]), which use either the theory-guided (top-down) approach, data-driven (bottom-up) approach or a combination of the two (Ercikan, Guo and He, 2020[22]; Guo and Ercikan, 2021a[23]; Guo et al., 2022[24]). For example, the normative threshold (Wise, 2017[16]) method is the most widely used data-driven approach, identifying the time threshold as x% of the average item response time, where x is a numerical value determined by the researchers. Responses with response time less than the normative threshold (NT) will then be classified as rapid guessing. While convenient, when using the NT method researchers need to validate the choice of x; that is, whether flagged item responses, using the selected threshold, lead to the chance score on average.

Another example is the mixture of log normal (MLN) distributions procedure, which is a parametric version of the visual inspection approach and a data-driven approach. It assumes that the item response time distribution is bimodal, where the lower mode distribution represents rapid responding, and the upper mode distribution indicates effortful responding. The lowest point between the two modes is then chosen to be the threshold. As with the NT method, users of the MLN method need to evaluate whether the flagged responses lead to the chance score or whether they reflect responses from highly competent students. Moreover, the MLN procedure cannot be used to identify a threshold in the absence of a bimodal distribution (Guo and Ercikan, 2021a[23]; Rios et al., 2017[25]; Rios and Guo, 2020[26]). In contrast, the cumulative probability method (Guo et al., 2016[27]; Guo and Ercikan, 2021b[28]) identifies the response time threshold at which the cumulative proportion correct rate begins to be consistently above the chance rate

for the studied item. By including the item correct rate (theory-guided) simultaneously with the timing distribution (data-driven), the cumulative probability procedure negates the need to separately validate whether flagged responses lead to a chance score on the studied item.

The above procedures produce item-level indicators (response time thresholds) to flag students' rapid responses on individual items on an assessment. Wise and colleagues proposed a test-level indicator for individual test takers, the response time effort (Wise, 2017[16]; Wise and Kong, 2005[29]; Hauser and Kingsbury, 2009[30]), to evaluate the overall effort on an assessment, which can be reframed as the rapid response rate for a test taker. The differential rapid responding measure was discussed with statistical significance in a large-scale assessment context (Guo and Ercikan, 2021a[23]; Rios and Guo, 2020[26]). Note that response time effort is equal to one minus the proportion of the number of rapidly responded items over the total number of items on the assessment.

Use of process data for improving assessment design and validity

Even when scores on an assessment rely on responses only (that is, scoring of final solutions to assessment tasks), process data can play an important role in improving assessment design and validating scoring meaning. This section focuses on discussing these uses of process data.

Improving assessment design

Response process data captured in digital assessments reflect how students interact with the assessment, which are traces of cognitive processes students use as they engage with assessments and "provide clues as to why students think the way they do, how they are learning, as well as reasons for any misunderstandings or gaps in knowledge and skills that may have developed along the way" (Pellegrino, 2020, p. 82[31]). Insights about response processes are particularly important in assessing complex constructs such as (collaborative) problem solving or communication, for which process is central to the construct and where examinees' sequences of actions provide more direct data on how they reason, collaborate and interact.

Analysing process data from assessments can inform task development with respect to the interfaces, representations, task features and scoring rules, as well as clarify any confusion and the measurement of the targeted construct. Assessment design typically relies on small-scale think aloud protocols (TAP) or cognitive labs to examine how assessment tasks function with students from targeted populations. Even though such studies are necessary and critical to provide valuable information, they tend to be based on small student samples. In technology-enhanced assessments, tasks can be administered to larger samples of students, which are more likely to reveal problems that may not surface with a small sample. For example, process data provide information about how examinees engage with different aspects of the assessment, how they navigate through the test, and if they go about solving problems and responding to questions as expected by content experts and TAP findings (DiBello et al., 2017[32]; He, Borgonovi and Paccagnella, 2021[33]; Nichols and Huff, 2017[34]). Often, unexpected behaviours are revealed in large-scale studies, such as rapid guessing behaviours (Wise, 2021[17]), which are unlikely to be observed in a cognitive lab study. Process data (particularly timing data) are also used in test assembly to manage the distribution of test time across sections and minimise test "speededness" (van der Linden, 2011[35]). In addition, process data can provide important insights about the functionality and use of tools included in assessments such as spell-check, read-aloud or text-to-speech functions, dictionaries, bilingual versions, and applications like a calculator (Wood et al., 2017[36]; Jiang et al., 2023[37]).

In interpreting process data for the purpose of improving task design, it is important to note that the same performance levels may be achieved through different behaviour patterns and test-taking behaviours may have different types of associations with performance for different populations. Test-taking behaviours are

associated with students' knowledge, skills and competencies targeted by the assessment, as well as with test-taking strategies and potentially their background and personal learning experiences, such as how they acquired the skills and how they prepared for the assessments (see Chapter 11 in this volume for more on score comparability). In the log data collected from a large-scale, high-stakes language test, distinct clusters were observed in test-taking behaviours regarding where and how long test takers spent their time on the test regardless of their test scores and testing modes (Guo, 2022[38]). In low-stakes assessments, test-taking behaviours are also confounded with students' test-taking motivation, strategies, cultural-language differences and other factors (Ercikan, Guo and He, 2020[22]; Guo and Ercikan, 2021a[23]; Wise, 2021[17]). Therefore, it is necessary to cross-validate different behaviour patterns in the process data with other sources (such as data from TAPs, survey questionnaires, socio-cultural context and results from different assessments) to evaluate possibilities of alternative interpretations, the impact of such differences on performance and ways to mediate or improve assessment designs, and to develop data analysis methods to consider biases introduced by sampling from diverse populations (especially heterogeneous test populations). In particular, assessments administered in multicultural contexts require validation of interpretation and use of process data and process indicators with the relevant diverse populations.

Validating score meaning

The importance of response processes in validating score meaning has been highlighted by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014[39]). In assessments of complex skills, process data can be used to investigate whether students' test-taking behaviours are aligned with the expected cognitive processes and to examine the extent to which items and tasks engage test takers in the intended ways and therefore can provide validity evidence (Ercikan et al., 2010[40]; Kane and Mislevy, 2017[14]; Yaneva et al., 2022[41]).

There is growing research demonstrating the usefulness of process data in validating performance on assessments as indicators of the targeted constructs. Based on experts' specified features that elicit particular components of knowledge, procedures and strategies, Carpenter, Just and Shell (1990[42]) developed production rules on the Raven's Progressive Matrices tasks (non-verbal geometric patterns presented in a matrix) to measure analytic intelligence. In addition to response error rates, they tracked students' eye movements and collected verbal protocol data. They found that the eye-fixation data supplemented by verbal protocols distinguished between a higher scorer and a lower scorer in the ability to induce abstract relations and to dynamically manage a large set of problem-solving tasks in working memory; that is, successful solutions were largely consistent with applying the expected or intended production rules in those data and thus reflected higher analytic intelligence. Also, using eye-tracking data collected from test takers responding to multiple-choice questions, Yaneva et al. (2022[41]) investigated alternative score interpretations by applying machine learning methods. The authors evaluated the prediction powers of a machine learning model on various combinations of features and found that different eye movement patterns were associated with correct/incorrect responses. Correct responses were associated with working from the item stem to the item options, spending more time on reading the problem carefully, and a more decisive selection of a response option aligned with the intended score interpretation.

Guo and colleagues (2018[43]) studied pause events (such as inter-key intervals and intra-word duration times) extracted from keystroke log timing data. They found informative and consistent features across different prompts in essay writing processes. These low-level features reflected students' writing fluency and showed added power in predicting writing performance. The findings were consistent with research on writing cognition which showed that keyboarding skills and composition fluency contribute to essay quality, particularly writing fundamentals (Deane, 2014[44]; Deane and Zhang, 2015[45]). Furthermore, using data mining techniques, Sinharay, Zhang and Deane (2019[46]) showed that process features from writing keystroke logs predicted essay scores nearly as well as their natural language processing (NLP) features, which provided validity evidence for process features and which in turn can be used to validate scores. Similarly, Greiff and colleagues (2016[13]) found that process indicators such as time on task, non-

interfering observation (observing how the problem environment behaved without any interference/actions) and intervention frequency with the test environment created from computer log files on some dynamic tasks and vary-one-thing-at-a-time (VOTAT) problem-solving strategies predicted performance, and thus enhanced score meaning.

Process data can also be used to provide information on test-taking strategies to offer insights on how test takers engaged with the assessment and the targeted assessment inferences and claims. For example, Greiff and colleagues (2016[13]) showed that time on task derived from complex problem-solving process data and performance had a complex relationship and followed an inverted U-shape relationship. Students who spent too little or too much time on the complex tasks showed overall poor performance on average. Intuitively, too little time may indicate that students were not engaged or had not fully understood the requirements of the tasks, while too much time spent might indicate exerting effort but having difficulty in solving the task, idling without engaging with the tasks or some other unexpected behaviours. He and colleagues (2021[33]) used process data to identify behavioural patterns in problem solving action sequences on interactive tasks (such as navigating through websites to search for a job) in a technology-rich environment. A process indicator was created that measured the distance or (dis)similarity between test takers' action sequences and the optimal strategies that content experts identified as the most efficient solution paths. Their results showed that test takers who followed optimal strategies were likely to obtain high scores, and thus supported score validity.

Use of process data as evidence of the construct for augmenting score creation

The second key use of process data is as evidence of the construct for augmenting the scoring of responses, which is particularly important for scoring interactive and simulation-based tasks where response processes are key aspects of assessing the targeted construct. These uses encompass their inclusion in the measurement model along with response data (Bennett et al., 2007[47]; Levy, 2020[48]); see also Chapter 8 of this volume.

Process data have been used in student profiles to augment performance. For example, using an Intelligent Tutoring System (ITS) for science learning, Betty's Brain, Biswas and co-authors (2015[49]) presented an overview of research on process data. This ITS uses the learning-by-teaching paradigm. Through studying students' actions and navigation behaviours with the system, video data and other process data, and utilising hidden Markov models, sequence mining and statistical analysis, the researchers were able to differentiate successful and unsuccessful performances and identify the sources of failure (such as lack of pre-knowledge). These results helped to reconceptualise and improve the system design and produce process indicators to cluster students into different profiles (such as frequent researchers and careful editors, strategic experimenters, confused guessers, disengaged, and engaged and efficient learners), which in turn helped to provide more informed scaffolding decisions to improve performance and student learning. Elsewhere, Guo and colleagues (2022[24]) applied several machine learning techniques such as autoencoder, unsupervised, semi-supervised and active learning on students' sequential process data such as response, timing and tool use sequences on a large-scale assessment to provide a holistic view of students' entire test-taking processes. Without sequential information, an isolated action or event is open to multiple interpretations – for example, rapid guessing behaviours at the beginning of the test may indicate low test motivation while rapid guessing behaviour at the end of the test suggests test speededness. In their study, students were grouped into profiles such as unengaged group, low engagement group, tool play and mixed test-taking strategy group, struggling group with high tool use, high performing with regulated time and tool use, etc. Such profiles helped to contextualise students' performance, augment score reporting and generate actionable feedback for educators.

Previous research examined incorporating process data into multidimensional latent models on digital-based assessments. Response time-based indicators, speed or response time latency measure how fast

test takers respond to items and are often jointly modelled with latent ability in a hierarchical model to investigate the speed-and-accuracy trade-off (van der Linden, 2007[50]; van der Linden and Fox, 2016[51]) and to detect aberrant response behaviours (van der Linden and Guo, 2008[52]). The joint model accounts for individual differences in speed when estimating ability and thus improves estimation accuracy (van der Linden, 2009[53]). Pohl and colleagues (2021[54]) extended the joint model to include the item omission propensity and speed-related process data (i.e. how often and fast they omitted items), in addition to latent ability and speed, to evaluate how process data-based indicators help to produce different profiles that reflect different aspects of performance. As Pohl and colleagues discussed, test-taking behaviours are not a nuisance factor that may confound measurement but an aspect providing important information on how test takers approach tasks. Test-taking behaviours have different relationships with performance for test takers with different backgrounds and learning experiences, which may lead to an unfair comparison of performance scores and difficulty in score interpretability. Therefore, researchers may experiment with different score reporting approaches such as reporting performance with supplementary profiles, including test-taking behaviours reflected in process data, to evaluate whether it can increase contextualisation, transparency and the valid use of assessment results.

From studies using process features, states and indicators generated from innovative assessments, new indicators based on process data have emerged to measure test takers' problem-solving processes as well as their knowledge. In particular, process data can be leveraged in technology-rich assessment environments to provide evidence of construct-relevant variables for complex constructs largely defined by processes. For example, using log files collected from an online learning system, Gobert et al. (2013[55]) applied data mining techniques on interaction features to replicate human judgement and measure whether students demonstrated science inquiry skills, which may provide real-time, automated support and feedback to students. Using simulation-based data from Harvard's Virtual Performance Assessments (VPA), Scalise (2017[56]) applied Bayes Nets to score rich patterns in the log data and then integrated that information into multidimensional Item Response Theory (IRT) models (e.g. mIRT-Bayes) to produce scores and yield inferences (see Box 8.1 in Chapter 8 of this volume for a more detailed discussion).

Andrews and colleagues (2017[57]) used process data and performance outcomes collected from about 500 pairs of test takers on a simulation-based task to analyse interaction patterns. Interaction patterns in chat box messages were annotated manually as categorical data and a multidimensional IRT model was used to analyse test takers' propensities toward different interaction patterns. Their results showed that different interaction patterns related to different performance outcomes and helped to conceptualise collaborative skills for assessment research. Similarly, Johnson and Liu (2022[58]) also developed a joint model that could simultaneously consider item response and process data, for example the onscreen calculator used on a mathematics assessment. A construct of the propensity of using a calculator was jointly estimated with the latent ability. Chapter 10 of this volume further showed that machine learning and Artificial Intelligence (AI) tools could be used to create features and/or indicators in collaborative tasks to augment score creation.

Note that in any assessment, the use of process data as evidence of the construct(s) critically depends on whether the assessment is designed to evoke the kind of processes that may be indicators of different levels of the targeted construct. Careful consideration of the use of process data in measurement models is necessary and their use should depend on whether their inclusion enhances the interpretation of assessment results.

Use of process data for group comparisons and fairness research

The third key use of process data is to examine differential engagement and possible construct-irrelevant variation in digitally-based assessments for different student groups (refer also to Chapter 11 in this volume for cross-cultural validity and comparability research using process data). Ercikan and Pellegrino (2017[59])

discussed this type of use for examining the comparability of response processes and patterns for students from different cultural groups, albeit with caution since some aspects of process data might not be directly comparable across different cultural and language groups (Ercikan, Guo and He, 2020[22]; Guo and Ercikan, 2021a[23]; Guo and Ercikan, 2021b[28]; He, Borgonovi and Paccagnella, 2021[33]).

In examining the equivalence of measurement for different groups, response process data can be used to examine whether students are interacting with items in expected ways and whether students engage with items similarly for the comparison groups (Ercikan, Guo and He, 2020[22]; Goldhammer et al., 2014[60]; Pohl, Ulitzsch and von Davier, 2021[54]; Yamamoto, Shin and Khorramdel, 2018[61]). These explorations of how examinees from different cultural and language backgrounds engage with items are particularly important when items involve interactivity and digital tools, such as graphing tools, dictionaries and search capabilities. Response process data such as item response time and the number of actions, which are captured in the Programme for International Student Assessment (PISA) (OECD, 2019[62]), for example, can be particularly important in examining how and the degree to which students from different cultural and language backgrounds interact with the assessment tasks, which in turn can inform inferences about the comparability of measurement across groups (Ercikan, Guo and He, 2020[22]; Goldhammer et al., 2017[63]).

As discussed earlier, response processes are functions of the targeted construct as well as test-taking strategies, the test taker's exposure to curriculum and instruction, their familiarity with the assessment technology, and their cultural and language backgrounds. When responding to a complex task, students need to have both the construct-relevant knowledge, skills and abilities (KSAs) that the assessment intends to measure as well as the construct-irrelevant KSAs (such as the ability to engage with digital interactive tasks effectively) that are necessary to understand, respond to and navigate through the assessments presented (Bennett et al., 2021[10]; Mislevy, 2019[64]; Sireci, 2021[65]; Sireci and Zenisky, 2006[66]). Studies have shown that relationships between item responses, response times or actions, and the construct may vary across different cultural and language groups (Ercikan, Guo and He, 2020[22]; Guo and Ercikan, 2021a[23]; He, Borgonovi and Paccagnella, 2021[33]; Pohl, Ulitzsch and von Davier, 2021[54]). Even for students with comparable performance on assessments in the same language, different test-taking behaviour patterns have been observed in students' keystrokes on essay writing tasks, response time sequences or action sequences in logfiles (Bennett et al., 2021[10]; Guo, 2022[38]; Guo et al., 2019[7]).

Differences in behaviours uncovered by process data may contain evidence on whether score comparability has been compromised, especially when such differences might hinder students' performance on assessments only for some language or cultural groups. For example, differences in input method editors can cause students in some language subgroups to devote much more time to typing. Digital platform features adapted and developed for one language subgroup may lead students in another language group to spend time figuring out task requirements. In these cases, students lose valuable time to work on other items which may adversely impact their performance. These insights highlight the importance of collecting rich process data to examine response processes and understand the potential source of process differences and their impact on the comparability of measurements and score meaning (Ercikan, Guo and He, 2020[22]; Kroehne, Deribo and Goldhammer, 2020[19]).

As mentioned in the previous section on validating score meaning, rapid guessing behaviour has clear relevance to the interpretation of group score comparisons. Rapid response behavioural differences may be attributed to various factors including differences in assessment context and language as well as cultural and motivational factors. In addition, because of cultural and language differences, the relationships between rapid response rates and performance may differ. Recently, a few process indices have been developed to identify differential response patterns for groups of test takers matched on performance on a test, such as differential response time and differential rapid response rate, to assess and evaluate the magnitude of differences in rapid response behaviours among different cultural and language student groups (Ercikan, Guo and He, 2020[22]; Ercikan and Por, 2020[67]; Guo and Ercikan, 2021a[23]; Kroehne, Deribo and Goldhammer, 2020[19]; Rios et al., 2017[25]; Rios and Guo, 2020[26]). These methodologies can

be applied to examine *differential engagement* in assessments by diverse student groups. It is important to note that, in addition to their possible impact on measurement comparability, differences in test-taking behaviours may provide important insights into how test takers approach and respond to assessment tasks. For example, Guo and Ercikan (2021a[23]) observed that student groups from different countries showed *differential rapid response rates*, even when there was ample testing time left; the lower-performing student groups showed a higher correlation between rapid response rates and performance.

Pohl and colleagues (2021[54]) showed differences in test-taking behaviours in terms of response time and item omission in different language groups. Differences in test takers' problem solving action sequences or writing state sequences were also observed among groups (Guo et al., 2019[7]; He, Borgonovi and Paccagnella, 2021[33]; Pohl, Ulitzsch and von Davier, 2021[54]), which prompted proposals of reporting test-taking behaviours on large-scale, low-stakes assessments as part of a performance portfolio across groups for fairer comparisons, a deeper understanding of performance and tailored interventions.

Group differences in performance can indicate proficiency level differences on the targeted construct. These differences could also be due to differences in construct-irrelevant factors such as familiarity with assessment technology or cultural relevance. Given their correlation with performance, process indicators like differential engagement, if reported as supplementary information to performance scores and survey results, may help educators and policy makers to identify the source of performance differences within and between groups, such as cultural norms/attitudes toward low-stakes assessments, lack of (pre-) knowledge, learning experiences and time management. When process data are appropriately integrated into measurement models, differential process indicators can provide insights to improve educational systems as has been done in learning systems (Biswas, Segedy and Bunchongchit, 2015[49]; Baker et al., 2019[3]).

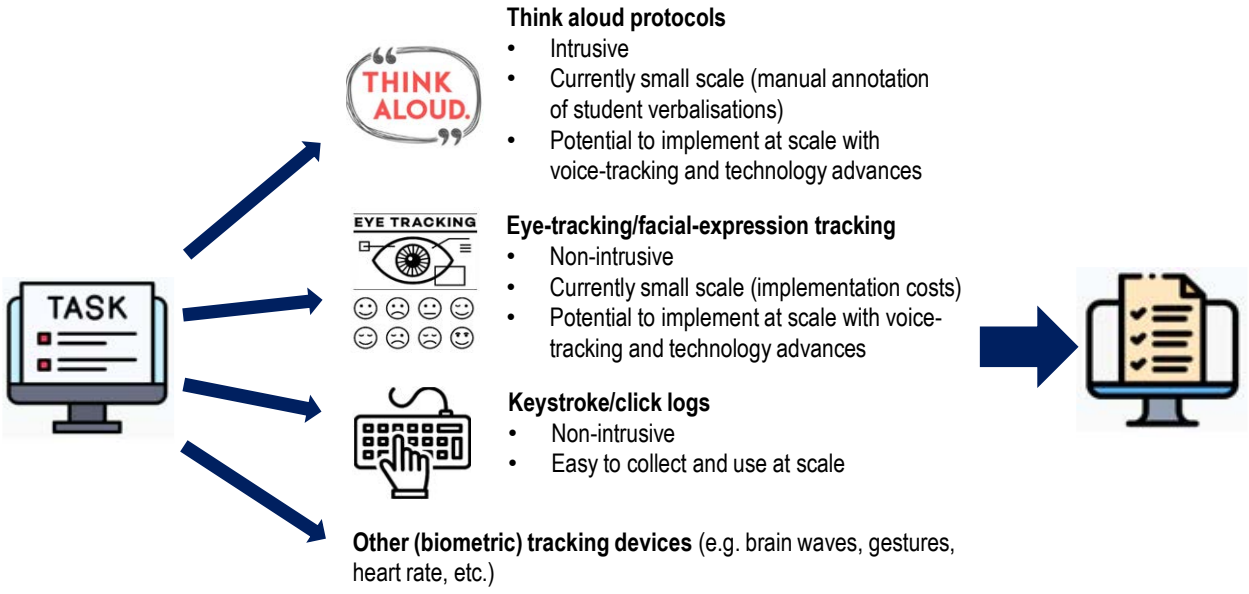
Conclusion

In this chapter, we have discussed different uses of process data in technology-rich assessments, from improving assessment design and validating score meaning to their uses as evidence of the targeted construct and for examining comparability and fairness for groups. We have reviewed relevant research that applied various research methodologies, from statistical and psychometric models to data mining and machine learning techniques for analysing process data for these purposes. Guided by cognitive theories, response process data including response time, action data and eye-tracking data collected during test administrations may help uncover differences in cognitive thinking processes and refine definitions of the target constructs, provide insights about sources of measurement variance, and consequently help examine score comparability and validity and improve assessment design. Findings from process data generated by technology-rich assessments can shed light on how students engage with tasks and solve problems and, eventually, help improve teaching and learning.

It is important to note two key considerations in using process data. First, process data as indicators of actions are only traces of students' thinking processes and require additional sources of information to help interpret these as evidence of response processes. The additional information that can help with interpretations can include data from think aloud protocols or cognitive labs (Bannert and Mengelkamp, 2008[68]; Benítez et al., 2016[69]; Ercikan et al., 2010[40]). The integration of data from multiple sources to support, validate and complement each other (for example, engagement is less likely to be an issue under watchful eyes in a cognitive lab) and to validate performance in uncovering cognitive thinking processes in mixed qualitative and quantitative methods have been used in some of the reviewed studies in this chapter (see also Table 10.1 in Chapter 10 as well as the discussion of developing the PISA 2025 Learning in the Digital World assessment in Chapter 9 of this volume, respectively). Process data such as think aloud and eye-tracking data can only be analysed at a small scale currently because of technical limitations and cost (see Figure 12.2 below). With advances in technologies and Artificial Intelligence, researchers

will be able to analyse multimodal data collected from technology-enriched assessments more efficiently and effectively at large scale.

Figure 12.2. Multimodal data capture to reflect thinking processes



Another issue to consider in using process data is the critical dependence of response processes on personal qualities and cultural and educational contexts. For example, the sequences and speed of actions in responding to assessment tasks depend on many individual factors such as anxiety levels, pre-knowledge, and learning and testing experiences (besides the actual targeted abilities). These highlight the limitations of interpreting response processes uniformly across student populations. One strategy that can help with such interpretations is to design tasks and task features that guide engagement with tasks in ways that support the interpretation of process data in the intended ways. As highlighted in earlier parts of this chapter, to take full advantage of technology-rich assessments tasks need to be developed and designed in such ways that their characteristics, features and directives can elicit the targeted cognitive processes and problem-solving behaviours so that process indicators can be captured from process data and in turn used as empirical evidence for task design, score creation and validation, and score comparability.

References

AERA, APA, NCME (2014), *The Standards for Educational and Psychological Testing*, American Psychological Association, Washington, D.C. [39]

Andrews, J. et al. (2017), "Modeling collaborative interaction patterns in a simulation-based task", *Journal of Educational Measurement*, Vol. 54/1, pp. 54-69, https://doi.org/10.1111/jedm.12132. [57]

Baker, R. et al. (2019), "Culture in computer-based learning systems: Challenges and opportunities", *Computer-Based Learning in Context*, Vol. 1/1, pp. 1-13, https://doi.org/10.5281/zenodo.4057223. [3]

Bannert, M. and C. Mengelkamp (2008), "Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning?", *Metacognition and Learning*, Vol. 3/1, pp. 39-58, https://doi.org/10.1007/s11409-007-9009-6. [68]

Benítez, I. et al. (2016), "Using mixed methods to interpret differential item functioning", *Applied Measurement in Education*, Vol. 29/1, pp. 1-16, https://doi.org/10.1080/08957347.2015.1102915. [69]

Bennett, R. et al. (2007), *Problem Solving in Technology-Rich Environments: A Report from the NAEP Technology-Based Assessment Project*, U.S. Department of Education, National Center for Education Statistics, Washington, D.C. [47]

Bennett, R. et al. (2021), "Are there distinctive profiles in examinee essay-writing processes?", *Educational Measurement: Issues and Practice*, pp. 1-15, https://doi.org/10.1111/emip.12469. [10]

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. [49]

Buchholz, J., M. Cignetti and M. Piacentini (2022), "Developing measures of engagement in PISA", *OECD Education Working Papers*, https://doi.org/10.1787/19939019. [15]

Carpenter, P., M. Just and P. Shell (1990), "What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test", *Psychological Review*, Vol. 97/3, pp. 404-431, https://doi.org/10.1037/0033-295x.97.3.404. [42]

Deane, P. (2014), "Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks", *ETS Research Report Series*, Vol. 2014/1, pp. 1-23, https://doi.org/10.1002/ets2.12002. [44]

Deane, P. and M. Zhang (2015), "Exploring the feasibility of using writing process features to assess text production skills", *ETS Research Report Series*, Vol. 2015/2, pp. 1-16, https://doi.org/10.1002/ets2.12071. [45]

DeMars, C. and S. Wise (2010), "Can differential rapid-guessing behavior lead to differential item functioning?", *International Journal of Testing*, Vol. 10/3, pp. 207-229, https://doi.org/10.1080/15305058.2010.496347. [18]

DiBello, L. et al. (2017), "The contribution of student response processes to validity analyses for instructionally supportive assessments", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York. [32]

Ercikan, K. et al. (2010), "Application of think loud protocols for examining and confirming sources of differential item functioning identified by expert reviews", *Educational Measurement: Issues and Practice*, Vol. 29/2, pp. 24-35, https://doi.org/10.1111/j.1745-3992.2010.00173.x. [40]

Ercikan, K., H. Guo and Q. He (2020), "Use of response process data to inform group comparisons and fairness research", *Educational Assessment*, Vol. 25/3, pp. 179-197, https://doi.org/10.1080/10627197.2020.1804353. [22]

Ercikan, K. and J. Pellegrino (eds.) (2017), *Validation of Score Meaning in the Next Generation of Assessments: The Use of Response Processes*, Routledge, New York. [59]

Ercikan, K. and H. Por (2020), "Comparability in multilingual and multicultrual assessment contexts", in Berman, A., E. Haertel and J. Pellegrino (eds.), *Comparability in Large-Scale Assessment: Issues and Recommendations*, National Academy of Education, Washington, D.C. [67]

Fratamico, L. et al. (2017), "Applying a framework for student modeling in exploratory learning environments: Comparing data representation granularity to handle environment complexity", *International Journal of Artificial Intelligence in Education*, Vol. 27/2, pp. 320-352, https://doi.org/10.1007/s40593-016-0131-y. [11]

Gobert, J. et al. (2013), "From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining", *Journal of the Learning Sciences*, Vol. 22/4, pp. 521-563, https://doi.org/10.1080/10508406.2013.837391. [55]

Goldhammer, F. et al. (2021), "From byproduct to design factor: On validating the interpretation of process indicators based on log data", *Large-Scale Assessments in Education*, Vol. 9/1, https://doi.org/10.1186/s40536-021-00113-5. [6]

Goldhammer, F. et al. (2017), "Relating product data to process data from computer-based competency assessment", in Leutner, D. et al. (eds.), *Competence Assessment in Education: Research, Models and Instruments*, Springer, Cham, https://doi.org/10.1007/978-3-319-50030-0_24. [63]

Goldhammer, F. et al. (2014), "The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment.", *Journal of Educational Psychology*, Vol. 106/3, pp. 608-626, https://doi.org/10.1037/a0034716. [60]

Greiff, S. et al. (2016), "Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files", *Computers in Human Behavior*, Vol. 61, pp. 36-46, https://doi.org/10.1016/j.chb.2016.02.095. [13]

Guo, H. (2022), "How did students engage with a remote educational assessment? A case study", *Educational Measurement: Issues and Practice*, Vol. 41/3, pp. 58-68, https://doi.org/10.1111/emip.12476. [38]

Guo, H. et al. (2018), "Modeling basic writing processes from keystroke logs", *Journal of Educational Measurement*, Vol. 55/2, pp. 194-216, https://doi.org/10.1111/jedm.12172. [43]

Guo, H. and K. Ercikan (2021b), "Comparing test-taking behaviors of English language learners (ELLs) to non-ELL students: Use of response time in measurement comparability research", *ETS Research Report Series* 1, pp. 1-15, https://doi.org/10.1002/ets2.12340. [28]

Guo, H. and K. Ercikan (2021a), "Differential rapid responding across language and cultural groups", *Educational Research and Evaluation*, Vol. 26/5-6, pp. 302-327, https://doi.org/10.1080/13803611.2021.1963941. [23]

Guo, H. et al. (2016), "A new procedure for detection of students' rapid guessing responses using response time", *Applied Measurement in Education*, Vol. 29/3, pp. 173-183, https://doi.org/10.1080/08957347.2016.1171766. [27]

Guo, H. et al. (2022), "Influence of selected-response format variants on test characteristics and test-taking effort: An empirical study", *ETS Research Report Series*, Vol. 2022/1, pp. 1-20, https://doi.org/10.1002/ets2.12345. [24]

Guo, H. et al. (2020), "Effects of scenario-based assessment on students' writing processes", *Journal of Educational Data Mining*, Vol. 12, https://doi.org/10.5281/zenodo.3911797. [8]

Guo, H. et al. (2019), "Writing process differences in subgroups reflected in keystroke logs", *Journal of Educational and Behavioral Statistics*, Vol. 44/5, pp. 571-596, https://doi.org/10.3102/1076998619856590. [7]

Hauser, C. and G. Kingsbury (2009), "Individual score validity in a modest-stakes adaptive educational testing setting", *Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego*. [30]

He, Q., F. Borgonovi and M. Paccagnella (2021), "Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks", *Computers & Education*, Vol. 166, https://doi.org/10.1016/j.compedu.2021.104170. [33]

Jiang, Y. et al. (2023), "Using sequence mining to study students' calculator use, problem solving, and mathematics achievement in the National Assessment of Educational Progress (NAEP)", *Computers & Education*, Vol. 193, https://doi.org/10.1016/j.compedu.2022.104680. [37]

Johnson, M. and X. Liu (2022), "Psychometric considerations for the joint modeling of response and process data", *Paper presented at the 2022 IMPS International Meeting of the Psychometric Society*. [58]

Kane, M. and R. Mislevy (2017), "Validating score interpretations based on response processes", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning for the Next Generation of Assessments*, Routledge, New York, https://doi.org/10.4324/9781315708591. [14]

Kroehne, U., T. Deribo and F. Goldhammer (2020), "Rapid guessing rates across administration mode and test setting.", *Psychological Test and Assessment Modeling*, Vol. 62/2, pp. 147-177, https://doi.org/10.25656/01:23630. [19]

Kroehne, U. and F. Goldhammer (2018), "How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items", *Behaviormetrika*, Vol. 45/2, pp. 527-563, https://doi.org/10.1007/s41237-018-0063-y. [5]

Levy, R. (2020), "Implications of considering response process data for greater and lesser psychometrics", *Educational Assessment*, Vol. 25/3, pp. 218-235, https://doi.org/10.1080/10627197.2020.1804352. [48]

McNamara, N. et al. (2011), "Citizenship attributes as the basis for intergroup differentiation: Implicit and explicit intergroup evaluations", *Journal of Community and Applied Social Psychology*, Vol. 21/3, pp. 243-254, https://doi.org/10.1002/casp.1090. [1]

Mislevy, R. (2019), "Advances in measurement and cognition", *The Annals of the American Academy of Political and Social Science*, Vol. 683/1, pp. 164–182. [64]

Nichols, P. and K. Huff (2017), "Assessments of complex thinking", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning in Next Generation Assessments*, Routledge, New York. [34]

OECD (2019), *PISA 2018 Results*, https://www.oecd-ilibrary.org/education/pisa_19963777. [62]

Pellegrino, J. (2020), "Important considerations for assessment to function in the service of education", *Educational Measurement: Issues and Practice*, Vol. 39/3, pp. 81-85, https://doi.org/10.1111/emip.12372. [31]

Perez, S. et al. (2017), "Identifying productive inquiry in virtual labs using sequence mining", in André, E. et al. (eds.), *Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science*, Springer, Cham, https://doi.org/10.1007/978-3-319-61425-0_24. [12]

Pohl, S., E. Ulitzsch and M. von Davier (2021), "Reframing rankings in educational assessments", *Science*, Vol. 372/6540, pp. 338-340, https://doi.org/10.1126/science.abd3300. [54]

Popp, E., K. Tuzinski and M. Fetzer (2015), "Actor or avatar? Considerations in selecting appropriate formats for assessment content", in Drasgow, F. (ed.), *Technology and Testing: Improving Educational and Psychological Measurement*, Routledge, New York, https://doi.org/10.4324/9781315871493-4. [2]

Rios, J. and H. Guo (2020), "Can culture be a salient predictor of test-taking engagement? An analysis of differential non-effortful responding on an international college-level assessment of critical thinking", *Applied Measurement in Education*, Vol. 33/4, pp. 263-279, https://doi.org/10.1080/08957347.2020.1789141. [26]

Rios, J. et al. (2017), "Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not?", *International Journal of Testing*, Vol. 17/1, pp. 74-104, https://doi.org/10.1080/15305058.2016.1231193. [25]

Roll, I. et al. (2005), "Modeling students' metacognitive errors in two Intelligent Tutoring Systems", in Ardissono, L., P. Brna and A. Mitrovic (eds.), *User Modeling 2005, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, https://doi.org/10.1007/11527886_48. [21]

Scalise, K. (2017), "Hybrid measurement models for technology-enhanced assessments through mIRT-bayes", *International Journal of Statistics and Probability*, Vol. 6/3, pp. 168-182, https://doi.org/10.5539/ijsp.v6n3p168. [56]

Schnipke, D. and D. Scrams (2002), "Exploring issues of examinee behavior: Insights gained from response-time analyses", in *Computer-Based Testing: Building the Foundation for Future Assessments*, Lawrence Erlbaum, Mahwah, https://doi.org/10.4324/9781410612250-20. [20]

Sinharay, S., M. Zhang and P. Deane (2019), "Prediction of essay scores from writing process and product features using data mining methods", *Applied Measurement in Education*, Vol. 32/2, pp. 116-137, https://doi.org/10.1080/08957347.2019.1577245. [46]

Sireci, S. (2021), "How psychometricians can affect educational assessment policy: A call to action", *Presentation at the International Test Commission (ITC) Colloquium*. [65]

Sireci, S. and A. Zenisky (2006), "Innovative item formats in computer-based testing: In pursuit of improved construct representation", in Haladyna, T. and S. Downing (eds.), *Handbook of Test Development*, Erlbaum, Mahwah. [66]

Sottilare, R. et al. (2013), *Design Recommendations for Intelligent Tutoring Systems - Volume 1 Learner Modeling*, US Army Research Laboratory, Adelphi. [4]

van der Linden, W. (2011), "Test design and speededness", *Journal of Educational Measurement*, Vol. 48/1, pp. 44-60, https://doi.org/10.1111/j.1745-3984.2010.00130.x. [35]

van der Linden, W. (2009), "Conceptual issues in response-time modeling", *Journal of Educational Measurement*, Vol. 46/3, pp. 247-272, https://doi.org/10.1111/j.1745-3984.2009.00080.x. [53]

van der Linden, W. (2007), "A hierarchical framework for modeling speed and accuracy on test items", *Psychometrika*, Vol. 72/3, pp. 287-308, https://doi.org/10.1007/s11336-006-1478-z. [50]

van der Linden, W. and J. Fox (2016), "Joint hierarchical modeling of responses and response times", in van der Linden, W. (ed.), *Handbook of Item Response Theory*, CRC Press, New York, https://doi.org/10.1201/9781315374512. [51]

van der Linden, W. and F. Guo (2008), "Bayesian procedures for identifying aberrant response-time patterns in adaptive testing", *Psychometrika*, Vol. 73/3, pp. 365-384, https://doi.org/10.1007/s11336-007-9046-8. [52]

Wise, S. (2021), "Six insights regarding test-taking disengagement", *Educational Research and Evaluation*, Vol. 26/5-6, pp. 328-338, https://doi.org/10.1080/13803611.2021.1963942. [17]

Wise, S. (2017), "Rapid-guessing behavior: Its identification, interpretation, and implications", *Educational Measurement: Issues and Practice*, Vol. 36/4, pp. 52-61, https://doi.org/10.1111/emip.12165. [16]

Wise, S. and X. Kong (2005), "Response time effort: A new measure of examinee motivation in computer-based tests", *Applied Measurement in Education*, Vol. 18/2, pp. 163-183, https://doi.org/10.1207/s15324818ame1802_2. [29]

Wood, S. et al. (2017), "Does use of text-to-speech and related read-aloud tools improve reading comprehension for students with reading disabilities? A meta-analysis", *Journal of Learning Disabilities*, Vol. 51/1, pp. 73-84, https://doi.org/10.1177/0022219416688170. [36]

Yamamoto, K., H. Shin and L. Khorramdel (2018), "Multistage adaptive testing design in international large-scale assessments", *Educational Measurement: Issues and Practice*, Vol. 37/4, pp. 16-27, https://doi.org/10.1111/emip.12226. [61]

Yaneva, V. et al. (2022), "Assessing the validity of test scores using response process data from an eye-tracking study: A new approach", *Advances in Health Sciences Education: Theory and Practice*, Vol. 27/5, pp. 1401-1422, https://doi.org/10.1007/s10459-022-10107-9. [41]

Zhang, M., H. Guo and X. Liu (2021), "Using keystroke analytics to understand cognitive processes during writing", *Virtual presentation at the International Conference of Educational Data Mining*, https://educationaldatamining.org/EDM2021/virtual/static/pdf/EDM21_paper_167.pdf (accessed on 13 March 2023). [9]

13 A tale of two worlds: Machine learning approaches at the intersection with educational measurement

By Kathleen Scalise, Cassandra Malcom and Errol Kaylor

(University of Oregon)

Promising digital technology affordances have expanded rapidly in education, and advances in the volume and nature of evidence that can be generated through digital technologies are impressive. However, especially at scale, analytical approaches to accumulate such data and draw meaningful conclusions (inferences) remain a frontier that is hard to navigate. This chapter discusses how machine learning and artificial intelligence approaches rapidly emerging in educational contexts are intersecting in many ways with educational measurement and argues for the imperative of these different fields to learn from each other. The chapter suggests some main takeaways for each field for the valid use and interpretation of innovative educational assessments.

Introduction

This publication makes the case that large-scale assessments should move beyond what is easy to assess and develop new methods to make claims on students' development of complex competencies. It also argues that simulations and other interactive assessment experiences can supplement more traditional item formats and potentially provide the evidence we need to compare students on these competencies. Scholars in learning analytics (LA) have made tremendous progress in applying data mining (DM) and machine learning (ML) techniques to the streams of data generated from learning experiences in digital environments. It is thus tempting to imagine that all we need to innovate assessment is to adopt these new techniques, moving beyond established methods of psychometrics. However, things are not so straightforward because the bridge between learning analytics and educational measurement still needs to be built.

The gap that exists between the two fields can be explained by their principal goals. The goal of LA is often to describe how learners learn or to find ways to adapt and personalise learning content to individual learners – sometimes also called Learning Engineering. LA uses a variety of data-driven approaches to make predictions about how people will learn in specific situations (i.e. using certain tools or working on certain activities). On the other hand, educational measurement – sometimes called psychometrics – focuses on making defensible claims about students' achievement, abilities or engagement with learning. Both can be used to inform learning interventions. LA and educational measurement are not completely distinct worlds: if the evidence collected for LA is then used to generate metrics (i.e. to make claims about students or student groups), then educational measurement is involved *even if the end result is framed as a prediction rather than a score report*. Some researchers including Sclater (2014[1]) and Wilson and Scalise (2016[2]) have therefore begun to establish standards of practice in LA when educational measurement is involved.

Hidden challenges for accumulating complex evidence

LA developers as well as those working in educational measurement have been converging in recent years, at least in very basic ways, on some important topics at the interface of LA and measurement technology. However, distance remains because scholars across these fields have different scholarly preparation, discourse language, epistemologies, ontological commitments and pedagogical grammars. For instance, in LA, conclusions tend to hinge on a relative argument about which model better explains the data set and is therefore the better "predictor" of something that the data set is purported to represent. This makes a lot of sense for traditional machine learning fields that rely on manifest variables (i.e. something that can be directly measured or observed) and involve data mining of homogeneous data sets – for example, a computer vision application that identifies if there is a tiger in a given picture (the tiger being the manifested variable). Yet for latent traits, such as mathematics knowledge and mental health, manifest variables do not exist and therefore there remain questions about what the data set represents.

Given these differences, LA scholars often see the need for new content and technology affordances in educational measurement but are unfamiliar with addressing what we refer to as the "Levy challenges". Over several decades, the field of psychometrics has developed well-accepted procedures for important issues in educational measurement, which include calibration and estimation of overall claims, reliability and precision information, test form creation, linking and equating, adaptive administrations, evaluating assumptions, checking data-model fit, differential functioning and invariance. However, as argued by Levy (2012[3]) and others, newer "data analytic" techniques using machine learning cannot rely on the same well-fitting measurement models used in psychometrics to verify the quality of assessment evidence and establish validity – an issue we call the "Levy challenges" and introduce in Chapter 8 of this report. LA approaches currently lack the theory and methods as well as the operational infrastructure often needed, such as analytic programs and delivery platforms at scale, to address these challenges. What this means

is that the way we can and should interpret findings about learners changes considerably if methods of evidence accumulation cannot satisfy these challenges via well-accepted procedures.

It should also be noted that classifications by software, such as describing how learners learn by classifying them into categories, also involves indicators and inferences. There would be no way to classify the learners if there were no indicators, and no reason to classify if no inferences were to be made from these classifications. Inferences might be about what might help the learner learn, but regardless, such inferences are claims about the student.

Many therefore advocate for a "separate-but-equal" view to break out different types of assessments by purpose. This sounds promising, but here is the resulting dilemma: if assessments for different purposes are treated as "separate-but-equal," this makes it simpler to analyse and implies no need to make consistent claims based on learning analytics methods and claims based on educational measurement – but this compromises the utility of assessments. Inconsistent claims across assessments – even if those assessments have different purposes (e.g. to plan future learning vs. to summarise what has been learned) – are highly confusing to teachers, students, parents and policy makers. For example, if teachers are told with one analytic technique that their students are proficient in their science knowledge while another technique says the same students have large learning gaps, teachers will not know what to believe and may ultimately discount the use of evidence in their work.

One solution that some suggest is to advance the field of assessment design by administering complex technology-based tasks at scale to generate process data but refrain from using the data for making inferential claims in the reporting (instead leaving these rich data for secondary research). However, such tasks may then be quickly discounted for various reasons: for example, is it a legitimate use of resources to develop technology-rich tasks if the new evidence they generate is not then used for reporting? Is it a legitimate use of student, school and administrators time to sit them? If such tasks are needed to fully actualise the construct, how can leaving out the evidence be justified? If the data are to be used for research, should the research not happen first and then deployment at scale occur when evidence is ready for reporting?

Another possibility is to neglect the "Levy challenges" altogether and be content with new analytic and reporting approaches that haven't yet matured robust measurement procedures. But as discussed throughout this report and elsewhere, the validity issues associated with integrating complex data into measures can be substantial. Without the ability to calibrate and estimate overall score(s), generate reliability and precision information, conduct subgroup analyses, create test forms, and engage in linking and equating, are robust inferences from large-scale assessments possible? Adaptive administrations, evaluating assumptions across languages and cultures, checking data-model fit, investigating differential functioning, and establishing invariance in the context of hard-to-measure constructs and complex naturalistic tasks goes well beyond what is currently known (Scalise, 2012[4]).

The intersection of the two worlds

As explored in more detail in Chapter 8 of this report, possible approaches to accumulating complex pieces of evidence to make inferences about students in large-scale assessment hold some promise. These include developing extended measurement models that borrow strengths from both psychometrics and learning analytics, establishing multiple inferential grain sizes or iteratively developing exploratory models to ultimately reach a confirmatory one. When considering any approach to accumulating (and eventually, reporting) assessment data, it is important to be mindful of the intended purpose of the assessment and of how results will be used and what they might be interpreted to mean. Given the hidden challenges described above, in the rest of this chapter we discuss some things that the fields of LA and educational measurement might learn from each other to further innovation in assessment analytics – especially in the context of large-scale assessments.

What learning analytics can learn from educational measurement

Evidence accumulation is as important as evidence elicitation

We argue that perhaps the most important thing that LA needs to learn about educational measurement is that when techniques such as machine learning (ML) – and more complex types of machine learning that may go by the name artificial intelligence (AI) – are used to make inferences about learning, it is not enough to consider only the elicitation of bits of evidence. The bits may look enticing and seem applicable to a given area of an assessment framework, but in and of themselves, bits of evidence alone do not satisfy a measurement argument.

As described in the Introduction chapter of this report, the interpretation vertex of the "assessment triangle" is actually two things: 1) defensible *elicitation* of bits of evidence; and 2) defensible *accumulation* of this evidence to make an inference. Exactly how all the pieces together add up to satisfy measurement claims and make the intended inferences is key. This accumulation of the bits to make inferences is called *aggregation* of evidence. Both elicitation of evidence and aggregation of that evidence to make claims 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 (as discussed in Chapter 11 of this report). Before reporting, both the elicitation and aggregation of evidence should be transparent, justified and warranted.

There must also be additional care paid to ensuring that features used in a predictive model are rigorously supported by educational frameworks. Feature engineering in educational assessment contexts should seek to provide meaningful data across levels of student skill rather than focus only on high performing students.

Measurement requires a chain of evidentiary reasoning supported by principled design

When technology is used for the purpose of educational measurement, often a set of questions are asked to guide the development of valid assessment instruments – such as in Evidence-Centred Design (see the Introduction and Chapter 6 of this report for a more detailed description). These questions are phrased differently in different contexts but essentially boil down to the following:

- *What can we do?* In other words, what do technology affordances allow to collect evidence and what "bits" of evidence will be elicited?
- *What do we want to report from the evidence?* In other words, how will we aggregate or accumulate evidence and what are the larger inferences and claims that the "bits" of evidence go together to form so that results can be meaningfully reported?
- *What will be needed to make the connection between the bits of evidence elicited and what we want to aggregate and claim?* In other words, what is needed to make the connection between the first two bullets?

A clear path based on evidentiary reasoning needs to be established between the goals and objectives of measurement (e.g. an assessment framework) and how the evidence is used to make a claim (e.g. the reporting of assessment results). In this way, stakeholders in education are empowered to use and value the results. It is also important to be able to use data to build a solid validity argument, for example verifying that conclusions about students are consistent across tasks that are designed to measure the same set of skills and evaluating that the assessment measures these skills well across the ability distribution and across sub-populations.

However, in many applications of LA based on data-mining approaches, the three principled design questions above are often not asked from the outset. A common feature of these approaches is the idea of discovering results from the data (i.e. generative or exploratory approaches) as compared to building

from theory (i.e. confirmatory). No argument is stated *a priori* regarding the aggregation of evidence but one is determined afterwards based on what is interesting in the evidence – for example, based on patterns in the data that might provide insights into students' cognitive processes in problem solving (Zhai et al., 2020[5]). This type of data exploration can provide highly interesting insights for research but might not be the best approach for making claims about what students can or cannot do. For making a claim, the analytics must be based on a clearly transparent and defensible argument using data.

Prediction is not the same as measurement

Learning analytics often aims to fit the best set of clusters, networks or other structures to data via machine learning, then defend the results based on having chosen the best fitting among various structures. Results are often considered acceptable if the predictions based on these indices are better than other models that were fit and more accurate than random. Patterns over a set of interactions may be possible to consider in measurement models, as discussed in Chapter 8 of this report, but treating predictions for such complex data as if they were valid measurements of latent traits can be problematic as the validity argument is not sufficiently defined nor is the validity evidence accumulated.

Producing a "new" metric just because the market asks might not be best way forward

In policy discussions around large-scale assessments there is often an urgency to assess new constructs and generate new insights, because there are many unanswered questions about students' skills and there is often a sense of *déjà vu* when new reports come out. There are also expectations that applying ML approaches to big data from technology-enhanced tasks will fill all of our existing information gaps. The "market demand" is undoubtedly there but this demand should not be the only driver of decisions in assessment design. What we need is better metrics not just new metrics. It is important to be responsive to policy makers when defining what we should assess and what insights on learners we need, but these new insights must be accompanied by a solid evidence base built on a carefully constructed validity argument. Otherwise, there is the risk that relevant findings are quickly discounted by measurement experts because they are not sufficiently validated and do not meet the "Levy challenges" described earlier.

Some things educational measurement can learn from learning analytics

Embrace the value of naturalistic tasks

Perhaps the most important thing that educational measurement needs to learn from researchers in learning analytics is the use of the naturalistic task. If policy makers want to take advantage of the many affordances of information technology discussed throughout this report, then the measurement field must be able to support the collection and aggregation of evidence from more authentic tasks and more complex activities. In a departure from the standard, multiple-choice assessments of the 20th century, naturalistic tasks are becoming more prominent in educational measurement. Naturalistic tasks can cover a broad array of task designs. Some examples come from science tasks, but many other disciplines also have created naturalistic tasks such navigating through a park in collaborative problem solving or reading with a purpose in mind to gather information for a presentation. In science and many other others, examples include simulations such as students interacting with lab equipment. These tasks are often defined by going through a natural development process where developers create the task similar to a classroom lesson that is guided by content standards (Scalise and Clarke-Midura, 2018[6]) and are further defined by realistic science experiences for students. Embedding small goals of a similar nature, such as assessing inquiry skills, can be done within and across several tasks with as much standardisation as possible in order to use ML/AI techniques to accumulate evidence on student learning (see again Chapter 8 of this report for an example of such an approach).

An important point here is that employing the naturalistic task isn't only about improving the precision and accuracy of the metrics. Measurement scholars will often want to discard complexity in innovation if metrics from simpler item types and tasks are likely to give the same measured result – or one that is similar enough to draw the same or similar inferences. Historically, for instance, in many contexts it has been shown that selected responses can measure some constructs as well as constructed response formats. Yet having students construct unique responses and evaluating those responses with rubrics or a scoring engine is nonetheless a central part of many assessments. Elsewhere in this publication (Chapters 2-5 and 7, particularly), the argument is made that using naturalistic tasks and gathering information on processes improves our evidence base for complex constructs such as collaborative problem solving, when essentially those constructs are largely defined as processes (e.g. collaboration is a process). By contrast, if only response data is considered and no opportunity is given to engage in such a process in an authentic way, the validity argument is certainly affected.

Why else? One reason is the importance of the "signifying" role of assessments. As discussed in the Introduction chapter of this report, educational research has established that teachers, students and local and national policy makers take their cues about the goals for instruction and learning from the types of tasks found on state, national and international assessments. Therefore, what is assessed in areas such as science, mathematics, literacy, problem solving, collaboration and critical thinking, and how those constructs are assessed, often will end up being the focus of instruction. In this role of signifying, it is hence critical that assessments represent the forms of knowledge and competency and the kinds of learning experiences we want to emphasise in classrooms. If students are expected to achieve the complex, multidimensional proficiencies needed for the worlds of today and tomorrow, they should be able to demonstrate their proficiency doing so. This requires moving away from *measuring what is easy* to *measuring what matters*.

Engagement and the student experience are also important considerations. Embedding agency and relevancy in an assessment activity is likely to increase students' engagement and thus the likelihood of observing what students can do at the best of their capacity. If we really want to describe what students know and can do, then test effort is an extremely important assessment argument.

However, we caution assessment developers not to get stuck in the "cluster buster"; not everything has to be as chunky as a long task, which may take a lot of student time and introduce a lot of unique variance that is construct-irrelevant. 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. Hard-to-measure constructs are often supported by measuring what is proxy and easy. To support interpretations, combining data types that are more known and less known is likely to remain important for making defensible claims.

To improve naturalistic assessments and the quality of measurements they can produce, an iterative and collaborative process between content experts, assessment developers, students and teachers is necessary. These ideas closely align with the concept of "Learning Engineering" as well as the Assessment Triangle discussed in the Introduction chapter of this report, emphasising an iterative process of building affordances and optimising learning experiences. Assessment developers must understand both how teachers look to understand student performance on a given task as well as how students engage with the task environment. These discussions help prioritise task modification to encourage positive task interaction styles among students as well as focus data collection in areas of interest and concern to the teacher.

Consider ways to establish a spectrum of comparability for reporting claims

The assessment field is still poised on the precipice of what assessing competencies using naturalistic tasks means. Costs, versioning, assessment platforms and other practical considerations exist, especially in the context of assessments that are intended to be replicable and comparable. Once such investments are made, often the need to handle longitudinal data also will emerge, further complicating what needs to

be in the measurement paradigm. Therefore, measurement standards may need to find ways to allow entry points that are not as difficult to satisfy. This is the classic solution in other fields with large advances in technology affordances. Can there be tiers and spectrums in education? For example, can there be a comparability spectrum across different purposes for the use of assessment evidence?

This might take the form of co-habiting for a time and focusing on 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 students are able to achieve in terms of their outputs from a given set of designed tasks; this could be what problems they were able to solve (given enough information of this type is generated across tasks similar enough to elicit the same latent trait). If this can be done, then LA might be used to provide more descriptive diagnostics of strategies and processes that students follow 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 applied 21st Century competencies.

Another perspective on the comparability spectrum question may be to not rely so heavily on the perfect fit of each item and score category or each scored observation in innovative assessments. Rather, either patterns (such as those discussed in Chapter 8 of this report) or a factor of larger tolerances might be allowed for the individual observations, if conclusions across the observations would be essentially the same. The same might be concluded about testlets and independence of information, by treating the issue as a discount factor on precision – so not overcounting the information when observations are not entirely independent as it is generally the case in extended naturalistic tasks.

It is hard to say what might be found with further research to simplify the intersection of fields. However, it is well known in educational measurement that no assessment of latent variables ever includes single observations. So less time might be spent in research looking at how individual observations vary, and rather researchers might look at if the inferences over the set vary substantially. Then the focus could be on the comparability of claims made across many items or many observations. However, the analytic mechanisms for how to approach these factors remain to be worked out.

What is at stake?

What is at stake for students may be no less than the development of broad transferable skills and knowledge. Some may believe it is possible to build such skills without assessment and without advanced digital technologies – arguing that such skills were necessary to achieve the accomplishments of earlier times before digital technologies were available. But of course, modern digital affordances are not only useful, as described in several other chapters of this report, but they are also today an expectation in the everyday lives of students. In many cases we don't yet understand the extent to which the broader transferable skills and knowledge that are the target constructs of complex assessments can be elicited across different digital tools – for instance, is self-regulated learning the same latent trait when applied in computational thinking contexts as when applied in reading literacy contexts? Will the knowledge-building tools that we might include in digital assessments to elicit evidence of such skills in turn cause them to manifest differently? What is the interaction of these skills with domain knowledge and domain tools? Such domain specifics no doubt impact the use and therefore performance of any such skills in an assessment context. But learning more about the common strategies students might engage as well as the differences in student behaviours that might arise between applications in digital contexts is important. To the extent there are commonalities, it seems key to understand them.

Conclusion

This chapter has discussed some important ideas that different fields approaching educational assessments might consider. However, an important audience for this report is policy makers. One message to policy makers from this chapter is that it is possible to undermine your objectives by overclaiming from emerging assessments. For instance, if a paramount concern in innovative assessment is ensuring equity and fairness as part of the validity argument, some of the issues discussed in prior chapters will likely require a softening of claims when reporting results from innovative assessments. An example is that policy makers should not select the most enticing wording for the "short" description when reporting on a construct if the wording is inaccurate. Even if this might make the innovative assessment seem more marketable, it will be a problem in the end if the "short" description does not match the claims it seems to be making with high quality evidence.

This will run counter both to what some policy makers want and what the market may demand. The exciting potential of technology affordances and rich new data sets may make policy makers want to lean into making strong claims for very new constructs. A suggestion is to be disciplined and wait on reporting strong claims until the needed progress in measurement science has been made – but don't stop developing and implementing such assessments or it won't be possible to make the needed progress. So do pursue innovation goals or you will never get there, but be mindful of your claims in the meantime.

To summarise, important opportunities will be sacrificed even longer by not incorporating new possibilities from new techniques (i.e. LA) even though this may require considerable exploration and grounding in traditional fields (i.e. educational measurement). Alternatively, important evidentiary techniques for measurement may be lost altogether and need to be recovered with much effort in the future if no pathway is created forward into a modern world. So, we argue, crossover between these two worlds is needed now. Wrestling with these topics will be hard and likely provocative since all sides will not be able to proceed as they currently do. Proceeding without change is also undesirable if fields are to meaningfully inform each other and affordances are to be optimised. Growing organically without incorporating field overlaps will likely lead to many missteps for new analytical approaches and will compromise trust in what the interpretation and use of results can mean for education.

Today, it is too soon to say what solutions might emerge at the intersection of the fields discussed here, but approaches such as hybrid models (see Chapter 8 of this report) or co-habiting (see earlier in this chapter) may become sufficiently established to represent a way forward. Regardless, it would seem inevitable that the bridge to a shared future will mean some spectrum of comparability is needed in educational measurement and assessment.

References

Levy, R. (2012), "Psychometric advances, opportunities, and challenges for simulation-based assessment", *Invitational Research Symposium on Technology Enhanced Assessments, K-12 Center at ETS*, https://www.ets.org/Media/Research/pdf/session2-levy-paper-tea2012.pdf. [3]

Scalise, K. (2012), "Using technology to assess hard-to-measure constructs in the ccss and to expand accessibility", *Invitational Research Symposium on Technology Enhanced Assessments, K-12 Center at ETS*, https://www.ets.org/Media/Research/pdf/session1-scalise-paper-2012.pdf (accessed on 14 March 2023). [4]

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. [6]

Sclater, N. (2014), *Code of practice for learning analytics: A literature review of the ethical and legal issues*, Jisc, https://www.wojde.org/FileUpload/bs295854/File/07rp_54.pdf (accessed on 9 April 2023). [1]

Wilson, M. and K. Scalise (2016), "Learning analytics: Negotiating the intersection of measurement technology and information technology", in Spector, J. (ed.), *Learning, Design, and Technology: An International Compendium of Theory, Research, Practice and Policy*, Springer, Cham, https://doi.org/10.1007/978-3-319-17727-4_44-1. [2]

Zhai, X. et al. (2020), "From substitution to redefinition: A framework of machine learning-based science assessment", *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1430-1459, https://doi.org/10.1002/tea.21658. [5]

14 Conclusions and implications

By James W. Pellegrino

(University of Illinois Chicago)

Natalie Foster and Mario Piacentini

(OECD)

This chapter reviews the arguments and evidence presented across the preceding chapters in this report in support of the proposition that innovation in assessment is both desirable and feasible. Many advances have been made in the conceptualisation, design and interpretation of innovative assessments of the 21st Century competencies that are and will be required of an educated citizenry for the foreseeable future. The chapter considers the challenges in bringing such innovative assessments to scale including the assembly of the intellectual, fiscal and political capitals needed to support such an enterprise. An example from PISA serves as an existence proof of the possible.

Introduction

This report argues for the need to seriously examine current assessment practices and pursue significant innovation in educational assessment. Such innovation encompasses: 1) the types of educational outcomes we assess; 2) how to do so by capitalising on the affordances of technology to systematically design assessment situations that provide rich and meaningful sources of data; and 3) the considerations needed to ensure that those assessments are valid and cross-culturally comparable given various possible interpretive uses to guide educational practice and policy.

This concluding chapter is divided into two sets of comments. Part I briefly restates key concerns expressed in the Introduction chapter for purposes of summarising and reflecting on the content of Chapters 1-13 of this report. The focus is on relevant theory and capabilities we now have while also considering the landscape of where we need to travel to advance the theory, science and practice of educational assessment. The chapters also remind us that assessment development is an application of scientific knowledge within constraints dictated by context and circumstances of use. Assessment is a "design science" like engineering, drawing upon foundational knowledge in the cognitive and learning sciences and in the measurement and data sciences. Part II of this concluding chapter attempts to look ahead and considers implications for what needs to be done to advance this design science and ensure its utility for education well into this century.

Part I. Arguments and evidence regarding innovation in educational assessment

The Introduction chapter of this report set forth three main arguments for innovating assessments that are elaborated and addressed across the following thirteen chapters in this publication. The first argument is that educational policy and practice need to consider what is important to measure and better define the components of what are often complex constructs and the authentic contexts in which we engage them. In education we need to *measure what matters* not simply *what's easy to measure*. The second argument follows from the first: to assess constructs that matter we need to innovate the ways in which assessments are designed, including the technologies used to assist in this process, bearing in mind the goal of generating useful evidence about what students know and can do with respect to these complex constructs. The third argument follows from the first two: for the results of any such assessments to be useful to the intended audiences – be they teachers, administrators or policy makers – they must be valid, i.e. assess those competencies that they purport to measure and not others. Thus the evidence generated by innovative assessments must accurately reflect the complexity of the constructs assessed while taking into consideration the diversity of the individuals assessed as well as the intended uses of the information.

The process of collecting evidence to support inferences about what students know and can do represents a chain of reasoning from evidence about student competence that characterises all assessments. This process has been portrayed as a triad of three interconnected elements – the *assessment triangle* – whose vertices represent the three key elements underlying any assessment: 1) a model of student *cognition* and learning in the domain of the assessment; 2) a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and 3) an *interpretation* process for making sense of the evidence considering the assessment's purpose and intended interpretive use (Pellegrino, Chudowsky and Glaser, 2001[1]). These three elements may be explicit or implicit, but an assessment cannot be designed and implemented, nor evaluated for its validity, without consideration of each.

Argument 1: Measuring what matters – Cognition

Several of this report's chapters explicitly focus on the "What" of educational assessment – the key constructs that we should be interested in assessing, why those constructs are important and where we

stand with respect to assessing them given the current educational assessment landscape. The bulk of the argument across Chapters 1-4 of this report is that we should be focused on complex cognitive and socio-cognitive constructs, the labels for which fall under the broader heading of "21st Century competencies." The challenge is carefully defining what we mean by these constructs so we can develop tasks and situations where individuals exercise the requisite competencies allowing us to obtain evidence that is valid, interpretable and useful whether the intended use is at the classroom level to guide learning and instruction or in a large-scale educational monitoring context such as the OECD's Programme for International Student Assessment (PISA).

Multiple reasons are offered for the need to assess such complex competencies. Among them is the fact that assessments have value beyond the data they provide. They can send key policy and aspirational signals by illustrating the types of performance we want students to master and, in so doing, they can become a driver of innovation rather than an obstacle to it. Preparing for (and sometimes engaging in) such assessments should engage students in instructionally valuable activities and results from the assessments should provide instructionally meaningful information. The tasks that students encounter should tap "higher-order" cognitive skills that support transferable learning rather than primarily emphasising skills that tap rote learning and the use of basic procedures. While there is a necessary place for basic skills and procedural knowledge in educational assessment it must be balanced with attention to critical thinking and applications of knowledge to new contexts.

Chapter 1 of this report provides an important review of different conceptual schemes that have been proposed regarding the complex cognitive and socio-cognitive competencies of interest, noting their family resemblances. The chapter also notes that while there is broad agreement and a strong narrative on the need to develop so-called 21st Century competencies, translating this vision into educational practice requires that curriculum, pedagogy and practice are aligned. In that regard, given its signalling power, assessment can be a powerful driver of alignment. The chapter is also realistic about the challenge of assessing 21st Century competencies with respect to defining constructs and learning progressions, the role of knowledge, task design, scoring, evidence interpretation, reporting and validation.

The challenge for task designers is to identify suitable test situations that call for students to engage 21st Century competencies as well as develop environments that allow students to respond authentically and generate interpretable evidence about their ways of thinking and doing. Along these lines, Chapter 2 of this report reinforces the importance of developing and using more interactive, complex and authentic assessment tasks, and argues that the theory and research in the learning sciences can provide important directions on how to do this. The authors propose some design innovations to develop new assessments that are complementary to existing ones. These include using extended performance tasks that have "low floors and high ceilings"; situating new assessments in specific knowledge domains; including opportunities for exploration, discovery and invention; and embedding intelligent feedback and learning scaffolds. If assessments are able to engage the same type of cognitive and socio-emotional processes of well-designed, deep learning educational experiences, then time spent on assessment need not take time away from learning.

Chapter 3 of this report takes the argument about the construct issue a step further regarding the cognition element of the assessment triangle and its implications for the assessment development process. While many agree that "higher-order" 21st Century competencies are important, the authors argue that identifying a long list of competencies and creating single assessment instruments for each one is probably not the most productive way to go. Such competencies do not exist in a vacuum and they are often used in combination in situations of real-life learning and problem solving. The chapter offers a simple framework to orient decision making on what to assess, guided by focusing on students' capacity to solve important problems. The framework includes: 1) deciding the type of activity students will work on (e.g. finding and evaluating information to make a decision; understanding how something works; designing a new product or process; building and communicating an argument, etc.); 2) deciding the context/knowledge required

by the activity (disciplinary or cross-disciplinary knowledge); and 3) deciding whether students will work on their own or collaboratively.

Elaborating on the close links between complex competencies and domain-specific knowledge and skills, Chapter 4 of this report argues that assessments and educational experiences must be re-evaluated to provide students with opportunities to engage and practice the type of decision making and problem solving that practitioners in a given domain face in the real world, e.g. learning how to think and reason like a mathematician, a scientist, an historian, etc. What this means is that assessment problems should require learners to choose strategies and make decisions that reflect authentic situations but nonetheless be constrained by requiring the knowledge expected only of students at a particular level. To make this possible, the authors argue, the key is to have a good understanding of the decisions practitioners face and define learning trajectory levels that are appropriate for the individuals assessed (cognition vertex of the assessment triangle), in turn using this knowledge to inform the design of tasks and scoring methods. The chapter operationalises this approach for an assessment of complex problem solving in science and engineering.

Argument 2: Principled assessment design and technology – Linking cognition, observation and interpretation

While Chapters 1 through 4 of this report focus on aspects of the *cognition* element of the assessment triangle, Chapters 5 through 10 address various aspects of the *observation* and *interpretation* elements of the assessment triangle with an emphasis on how technology can be exploited through and within a principled design process to create robust assessments instruments.

Chapter 5 of this report emphasises the point that digital technologies significantly expand our assessment capabilities. The author notes that innovative technology-enhanced assessments have the potential to expand the competencies it is possible to assess, enhance the way that tasks are designed and presented, as well as generate new sources of data and methods for analysing such evidence. Technology also introduces the possibility of embedding assessments in learning environments capable of providing unobtrusive and more contextualised data on problem solving and decision-making processes to complement data from stand-alone assessment scenarios. Regardless of the context for data capture, the author underlines that the use of technology to assess complex constructs must be done following a coherent and principled design assessment process.

Chapter 6 reminds us that every new assessment needs to be guided by a theory of knowing and learning in the domain of interest and that anchoring the design of tasks and the evidence model to a well-defined theoretical framework is essential for generating valid inferences about performance – especially in the context of assessing complex constructs. The author emphasises that this process is an exercise in design science; one requiring close collaboration among potential users of the results, domain experts, psychometricians, software designers and UI/UX experts from the beginning of the design process. Evidence-Centred Design (ECD) provides an architecture to structure the process and products of such a collaborative design activity. The author notes that the main challenges in this process lie in interpreting the data that innovative and open tasks provide. We have much to learn about modelling the complex data associated with multidimensional and dynamic constructs including in situations where students can “learn” by interacting with the assessment environment. Some of these modelling and interpretive challenges are taken up in Chapters 8 and 13 of this report.

Chapter 7 of this report reminds us that the features of many educational assessments, particularly large-scale standardised tests, have been designed and formalised given various “practical” constraints (e.g. administration and scoring costs, time) and “technical” constraints (e.g. psychometric standards of reliability, validity, comparability and fairness). Today, in large part thanks to technological and data analytic advances, more is possible in terms of designing task formats, test features and sources of evidence for assessments. Notably, innovative assessments allow us to move beyond static tasks to

present individuals with interactive and immersive problems. Such situations can also be designed to be adaptive to the test taker, to include resources for on-the-fly learning and to capture process data in addition to product data.

Chapter 8 of this report addresses how innovative assessments might capitalise on these technology affordances to generate defensible measurement claims that allow us to make inferences about respondents or the groups they represent, given that typical psychometric models used operationally in large-scale assessment programmes do not easily or well incorporate such complex data. While newer "data analytic" techniques can handle large and complex data sets that include process data, they do not yet have mature machinery appropriate to meeting the challenges of making measurement and inferential claims. The chapter discusses new analytical methods where existing models in psychometrics and data-mining techniques borrow strength from each other directly. It exemplifies a mIRT-Bayes hybrid approach that integrates scores generated by Bayes nets into an Item Response Theory (IRT) model, generating sizable measurement precision gains. The author argues that these approaches exploit the suitability and flexibility of Bayes nets for describing construct-relevant patterns from process data in technology-enhanced tasks while preserving the robust statistical properties of latent variable methods.

Chapters 5-8 of this report collectively make the case that processes of principled assessment design can take advantage of the affordances of technology to expand the space of the possible for design and implementation of next-generation assessments, and describe how the cognition, observation and interpretation components of the assessment triangle can be linked together to enable the reasoning from evidence process that must accompany any valid assessment effort. Chapters 9 and 10 of this report then provide more concrete cases of the possible, illustrating some of what can be done with technology while presenting some of the challenges inherent in working in complex design and interpretive spaces.

Chapter 9 argues that as educational goals increasingly focus on students' capacity to learn, assessment should also enable and evaluate that capacity. Accordingly, assessment situations should invite students to engage in scenarios that can help elucidate the processes of learning. The chapter provides an example of designing feedback and resource affordances embedded in assessment/learning scenarios to serve two purposes: 1) to support authentic learning; and 2) to provide evidence about learning processes and how learners regulate their learning. The authors note that the design and use of resources introduces many challenges regarding the validity of inferences – perhaps most notably regarding the role of prior knowledge. Typically, the knowledge that learners bring to an assessment is the target construct of the assessment; in the example discussed in the chapter, the assessment is the context in which learning takes places. Interpreting learning activities in light of prior knowledge is a major undertaking, as learners' activities and strategy choices are contingent on their knowledge. A related challenge is that of generalisability and transfer of students' ability across resource types, learning opportunities and tasks. The authors conclude that inferences about the use of tools should combine top-down (justified by theory) and bottom-up (visible in data) arguments and evidence.

In Chapter 10 of this report, the authors turn to Intelligent Tutoring Systems (ITS) to illustrate how advances in technology and data analytics (e.g. natural language processing, speech recognition software, etc.) can enable the sorts of innovative designs argued for in previous chapters. ITS, the authors contend, exemplify how digital environments can already provide learners with dynamic learning tasks, interactivity and constant feedback loops – and hence innovative assessments have much to learn from them. With a number of examples from ITS, the chapter emphasises how artificial intelligence (AI)-based applications can support task design and scoring, from automating intelligent feedback tailored to the actions of examinees to producing indicators of learners' collaboration with others in open scenarios.

Argument 3: Ensuring valid assessments

Validity is the single most important property of any assessment, yet an assessment is never valid in and of itself: an assessment that may be valid for one interpretive use (e.g. a classroom teaching situation)

may be invalid for a very different interpretive use (e.g. a cross-national comparison), and vice versa. As such, an assessment's validity depends on arguments and evidence about its specific interpretive use. To be valid for a wide range of learners, assessments of complex constructs should *measure well* what they purport to measure, be *accurate* in evaluating students' abilities, and do so *reliably* across assessment contexts and scorers. They should also be *unbiased* and *accessible* and used in ways that support positive outcomes for students and educational systems. Principles associated with establishing validity are especially important and deserve careful attention and investigation for technology-rich assessments that target the measurement of complex performances.

Much of this report focuses on one of the most critical aspects of establishing the validity of next-generation assessments for 21st Century competencies: validity arguments and evidence derived through the application of a principled design process. For example, Chapter 6 discusses the key decisions that need to be considered and addressed at the beginning of an ECD process to guide development of valid assessments of complex constructs, starting with the definition of the construct(s) of interest. Evidence from the design process would then be complemented by various forms of empirical data on how the assessment performs, and the entire complex of evidence would constitute the elements of an assessment's validity argument. Chapters 11 and 12 of this report address particular issues of validity and comparability in large-scale, technology-rich assessments including methodologies and principles for examining validity issues throughout assessment design and once data have been collected.

Chapter 11 notes that complex constructs like creativity, critical thinking, problem solving or collaborative skills are characteristically shaped by cultural norms and expectations. As a result, challenges arise in balancing measurement validity with score comparability in multilingual or multicultural assessment contexts. Therefore assessment developers should consider construct equivalence, test equivalence and testing condition equivalence during the assessment design process. Use of digital assessments, especially for assessing complex skills, also necessitates evaluating students' digital literacy and examining potential biases against cultural subgroups in AI-based methodologies such as in test adaptivity, automated scoring and item generation engines.

Chapter 12 of this report expands on the very important point that process data, as mentioned in Chapters 7 and 8, can serve as important evidence regarding the processes of reasoning and problem solving that individuals employ when they work on complex assessment tasks irrespective of whether those tasks are stand-alone or are embedded in broader learning environments. Such process data can function in two ways related to assessment validity. Studies using process data in complex tasks have shown their value in validating assumptions about the cognitive constructs involved in assessment performance and as such they can constitute critical data during assessment design and initial validation efforts before assessments become fully operational. For tasks where prior validation of performance has been done, process data obtained during task execution may enrich score meaning and reporting and constitute a part of the interpretive process and evaluation of performance that goes beyond scores based solely on response accuracy. For example, differential engagement with an assessment task or situation is a potentially important index for both practitioners and policy makers. Students' performance on large-scale assessments may not be a pure reflection of what they actually know and can do because of differences in prior knowledge, cultural norms, familiarity with technologies, attitudes and differences in educational experiences.

Finally, one of the implications throughout this report regarding the design of complex tasks and performances is that the interpretation of the evidence provided will not be simple. Undoubtedly it will require models and interpretive schemes that go well beyond the psychometric models and methods that have been the mainstay of most large-scale assessments. Chapter 13 of this report discusses validity implications of using the results of "predictions" from data mining and machine learning methods in reporting, given that these analyses are not supported by validity evidence that is deemed central in educational measurement (e.g. on reliability and precision, check of data-model fit, differential functioning and invariance). The authors argue that there is an important and critical intersection emerging between

the fields of educational measurement and learning analytics, issues of vocabulary and definitions notwithstanding. These broad fields, which are really many fields, can meaningfully learn from each other when making claims or inferences about the complex constructs represented in innovative assessments. By engaging in solving the measurement and inferential issues that currently exist, both fields will likely advance the science and practice of educational assessment.

Part II. Innovative assessments: Progress made and the road ahead

No single assessment can evaluate all the forms of knowledge and skill that we value for students, nor can a single instrument meet all the goals and information needs held by parents, practitioners and policy makers. As argued in the Introduction chapter, we need coordinated systems of assessments in which different tools are used for different inferential and reporting purposes – for example, formative and summative, or diagnostic vs. large-scale monitoring. Such assessment tools would operate at different levels of the educational system from the classroom on up to school, district, state, national and/or international levels of application. Within and across these levels, all assessments should faithfully represent the constructs of interest and reflect and reinforce desired outcomes that arise from good instructional practices and effective learning processes.

As noted in the Introduction chapter, the following features define the elements of assessments that operate within and across such systems of assessment: 1) the assessment of higher-order cognitive skill; 2) high-fidelity assessment of critical abilities; 3) items that are instructionally sensitive and educationally valuable; and 4) assessments that are valid, reliable and fair. A major challenge is determining a way forward whereby we can create coherent systems of assessments that meet the goals we have for the educational system, satisfy the information needs of different stakeholders, and that align with these criteria. The chapters in this report reveal progress that has been made in conceptualising and operationalising critical aspects of the assessments needed within such systems. The report provides a vision of what next-generation assessments 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 along the way.

A journey of the type envisioned by this report's body of work cannot be undertaken nor will it succeed without an investment of multiple forms of capital. In the discussion that follows we consider three particular forms of capital that are needed and expand on why each is critical to the success of such an endeavour. They include intellectual capital, fiscal capital and political capital. Each is necessary but insufficient on its own – yet 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

The collective work described in this report illustrates that no single discipline or area of expertise will be sufficient to accomplish what needs to be done to innovate assessment. Advances to date reveal that next-generation assessment development is inherently a multidisciplinary enterprise: different communities of experts need to work together collaboratively to find solutions to the many conceptual and technical challenges already noted as well as those yet to be uncovered as part of the journey. 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 for different student groups, particularly those for complex skills, 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 assessment of creative thinking (OECD, 2022[2]) exemplifies comparability challenges related to assessing a complex construct across language and cultural groups (see Box 11.1 in Chapter 11 of this report for more). Systematic evaluations of measurement comparability can provide the basis for future assessments of complex skills.

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 by working with experts from different disciplines, machine learning and AI techniques can help researchers better understand and model learning processes (Kleinman et al., 2022[3]) and can assist content experts in efficiently and effectively annotating students' entire problem-solving processes at scale (Guo et al., 2022[4]). Work of this type is needed to supplement evidence derived from small-scale cognitive lab studies, advance learning science and have an impact on large-scale assessment.

At a pragmatic level, Schwartz and Arena (2013[5]) 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 Platform for Innovative Learning Assessments[1] (PILA) at the OECD, 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 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 AI, machine learning, statistics and psychometrics. Collaborative engagement with these concerns by learning scientists, data scientists, measurement experts, assessment designers, technology experts, experts in user interfaces and educational practitioners could yield a new discipline of Learning Assessment Engineering.

Fiscal capital

The development of assessments for application and use at any reasonable level of scale is a time consuming and costly enterprise, especially for innovative assessment of the types envisioned in this report. 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 National Assessment of Educational Progress (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

[1] https://pilaproject.org/.

programmes have moved to technology-based task presentation, data capture and reporting. Capitalising on many of the affordances of technology as described in this report has not been a distinct feature of those assessment programmes.

Developing and validating technology-rich assessment tasks and environments of the type advocated for in this report 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 competencies is the paucity of examples of assessment instruments of complex cognitive and socio-cognitive constructs, especially examples that have been built following systematic design principles 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. To advance the field of 21st Century innovative assessment, considerably larger capital investments need to be made of two types as argued below.

Substantial fiscal capital is required to assemble and support the multidisciplinary teams needed to conduct research and development supporting the creation of innovative next-generation assessments. The amount currently invested in multidisciplinary assessment research and development (R&D) work are but a tiny fraction of what is spent on more conventional large-scale assessment development and implementation. Neither government funding agencies, private foundations, testing companies nor governmental assessment agencies have been willing to make the systematic and sustained investments required. Funding at fiscal levels representing a small percentage of the total fiscal expenditures on educational assessment would make a significant difference in what could be done and the time to do so. Without sustained and increased investments in the types of work required it will prove difficult, if not impossible, to accumulate the knowledge required to solve the conceptual and technical problems that remain and generate the solutions required for valid and useful assessment of challenging constructs.

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 policy makers. 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 students' knowledge and skills are no longer seen as discrete and independent, assessing them requires examining the entire interactive 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 a transformation process for educational assessment in ways envisioned in this report.

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 outlined throughout this report. One of the biggest challenges in making change happen is that scale is needed to show what is possible. As noted earlier, 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 in turn have taught us much with respect to what is possible as well as where challenges remain.

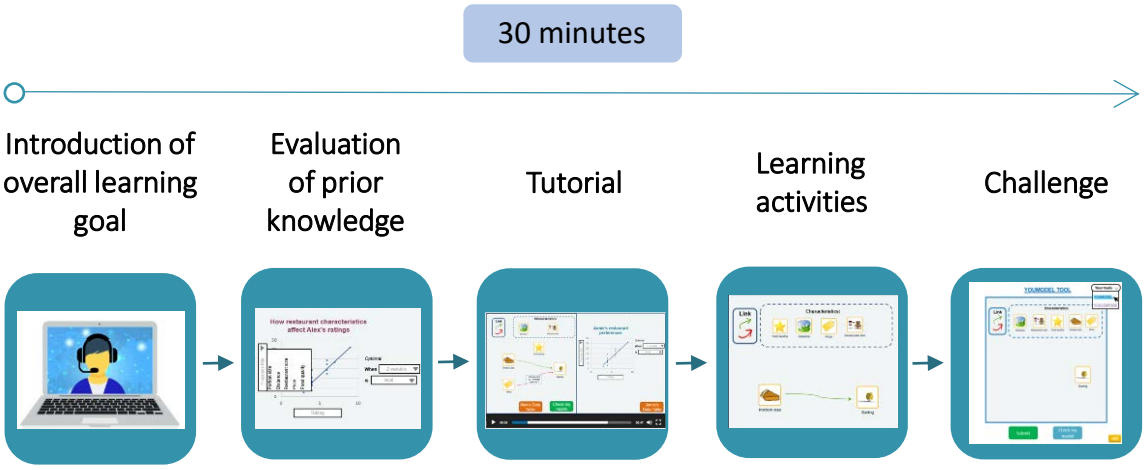
International assessments generally serve as tools for monitoring performance on contemporary disciplinary standards. As such these programmes make statements about what is valued globally and provide information about student proficiency at scale. They also illustrate an operational example of the pooling 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 programme in mathematics, reading and science, the 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.

In its 2025 cycle, PISA will include an assessment of Learning in the Digital World (LDW). 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 14.1).

Figure 14.1. Task sequence in the PISA 2025 Learning in the Digital World assessment



Source: Adapted from OECD (forthcoming[6]).

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 later, 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 unit, 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 also how they go about learning to do so.

These innovations represent responses to well-defined evidentiary needs. As further elaborated in Chapter 6 of this report, the assessment has been designed to provide responses to three interconnected questions: 1) what types of problems in the domain of computational design and modelling can students solve? 2) To what extent are they able to learn new concepts in this domain by solving sequences of connected, scaffolded tasks? And 3) to what extent is this learning supported by productive 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 multidimensional reports of student performance on this test including measures of: 1) students' 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) students' 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 in this chapter. 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 student groups.

This new PISA test represents only an initial foray into the enterprise of innovating assessments. As argued in this report, 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 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

Guo, H. et al. (2022), "Understanding students' test performance and engagement", *Invited session*. [4]

Kleinman, E. et al. (2022), "Analyzing students' problem-solving sequences", *Journal of Learning Analytics*, Vol. 9/2, pp. 1-23, https://doi.org/10.18608/jla.2022.7465. [3]

OECD (2022), *Thinking Outside the Box: The PISA 2022 Creative Thinking Assessment*, OECD Publishing, Paris, https://issuu.com/oecd.publishing/docs/thinking-outside-the-box (accessed on 16 April 2023). [2]

OECD (forthcoming), *PISA 2025 Learning in the Digital World Assessment Framework*. [6]

Pellegrino, J., N. Chudowsky and R. Glaser (eds.) (2001), *Knowing What Students Know*, National Academies Press, Washington, D.C., https://doi.org/10.17226/10019. [1]

Schwartz, D. and D. Arena (2013), *Measuring What Matters Most: Choice-Based Assessments for the Digital Age*, The MIT Press, Cambridge. [5]

Innovating Assessments to Measure and Support Complex Skills

Policy makers around the world recognise the importance of developing young people's 21st century skills like problem solving, creative thinking, self-regulation and collaboration. Many countries also include these skills as part of the intended learning outcomes of their education systems. To shift intention into practice, educational assessments need to better measure what matters. Innovative assessments are needed that combine conceptual, technological and methodological advances in educational measurement.

This report explores new approaches to measuring complex skills through a practical and applied assessment design lens, bringing together perspectives from leading experts to consider what we can learn from the learning sciences to define more authentic assessment experiences and expand the range of skills we are able to measure in both disciplinary and cross-disciplinary contexts of practice. The report also examines how technology can expand our possibilities for innovation, including the creation of more interactive and immersive problems and the generation of meaningful sources of potential evidence about students' proficiency. Finally, the report explores how we can make sense of the rich data captured in interactive digital environments using new analytical approaches, and how we can ensure the valid interpretation and use of results from innovative assessments.

PRINT ISBN 978-92-64-66443-2
PDF ISBN 978-92-64-37850-6

9 789264 664432