

For Official Use**English text only**

27 February 2023

**DIRECTORATE FOR EDUCATION AND SKILLS
PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT****Governing Board****THE USES OF PROCESS DATA IN PISA****55th meeting of the PISA Governing Board**

21-23 March 2023

Rome, Italy

The PGB is invited to:

- **NOTE** and **COMMENT** on the working paper.

Andreas Schleicher, Director for Education and Skills and Special Advisor on Education Policy to OECD's Secretary-General (andreas.schleicher@oecd.org).

JT03513136

The Uses of Process Data in PISA

Abstract

The advent of computer-based testing is making available more data to evaluate the performance of students. Using the data that are automatically logged by the digital platform, we can now describe the respondents' performances not only considering their response accuracy, but also their responding processes. The potential to collect a rich set of process data in large-scale assessments is opening the door to a wide range of possible uses of these data across the assessment cycle. From the definition of test domains all the way to scoring and reporting, process data are changing the way we conceive assessments and have the potential to improve assessment validity and fairness, construct measurement, and quality assurance processes. The present paper provides an outlook of the range of process data that can be captured in assessments at a small- and large-scale and presents a framework for a more systematic use of process data across the PISA cycle. For each phase of the cycle, we describe how process data can be leveraged to improve the quality of PISA by drawing examples from the literature and from the development of the PISA 2025 innovative assessment Learning in the Digital World. The paper then highlights the challenges arising from the collection and use of process data. These relate to their valid interpretations, their analysis, and associated ethical concerns. Perspectives for overcoming these challenges and taking full advantage of process data in PISA are discussed.

Table of contents

| | |
|--|----------|
| The Uses of Process Data in PISA | 2 |
| Abstract | 2 |
| <i>Table of contents</i> | 3 |
| Introduction | 4 |
| Section 1: What are process data? | 5 |
| Section 2: Using process data across the PISA cycle | 8 |
| Framework development | 8 |
| Item development and scoring | 12 |
| Test construction | 19 |
| Data adjudication and quality check | 20 |
| Development of proficiency measures | 21 |
| Dissemination | 22 |
| Section 3: Challenges to the use of process data in PISA | 24 |
| Validity challenges | 24 |
| Ethical challenges | 24 |
| Analysis challenges | 25 |
| Section 4: Conclusions | 27 |
| References | 28 |

FIGURES

| | |
|---|----|
| Figure 1. The uses of process data across the PISA cycle | 8 |
| Figure 2: Iterative design to use process data for assessment measurement | 13 |
| Figure 3: Example task from the PISA 2025 Learning in the Digital World prototype unit "I like that!" | 14 |
| Figure 4: Example student model for 'Computational and scientific inquiry practices' | 14 |
| Figure 5: Process of carrying out control of variable strategy to find a relationship between two variables | 16 |

TABLES

| | |
|--|----|
| Table 1: Types of process data | 6 |
| Table 2: Example evidence rule table for conducting controlled experiments | 16 |
| Table 3: Partial task model for conducting controlled experiments | 18 |

Introduction

1. Traditionally, large-scale assessments have focused on evaluating students' competences based on their ability to achieve correct responses in the test. However, results from these assessments provide only limited information on how students learn, perform and progress through complex tasks. In addition, such an approach makes it difficult to assess complex competences, including '21st century skills' such as collaboration or self-regulated learning, which are mostly defined by processes, rather than by outcomes.
2. The transition to technology-based assessments is introducing new opportunities to incorporate more open, interactive and engaging tasks that require students to engage in complex response processes. These opportunities come hand in hand with the possibility to digitally capture fine-grained data on students' actions throughout the tasks (such as students' mouse clicks), thus enabling inferences on students' thinking processes.
3. The potential to collect a rich set of process data in large-scale assessments is opening the door to a wide range of possible uses of these data across the assessment cycle (Provasnik, 2021^[1]). From the definition of test domains all the way to scoring and reporting, process data are changing the way we conceive assessments and have the potential to improve assessment validity and fairness, construct measurement, and quality assurance processes.
4. The Programme for International Student Assessment (PISA) has started to gather evidence on the processes followed by students as they work on interactive tasks, particularly in the Science and 2012 Creative Problem Solving domains. Further advances have the potential to strengthen the value of PISA data for policymakers and other education stakeholders by providing information not just on the questions students can answer to, but also on their capacity to engage in important strategic and monitoring processes. Moreover, process data offer opportunities to collect highly relevant diagnostic information that help interpreting performance scores and can inform interventions. For example, using process data we can make inference on whether students fail to understand a text because they lack basic reading fluency, or on whether they cannot provide a correct response to a mathematics problem because they make procedural mistakes.
5. The present paper offers an overview and framework for a more systematic use of process data in PISA. After providing an outlook of the range of process data that can be captured in assessments at a small and large-scale (including, but not limited to computer logfiles) (Section 1), the paper presents a framework for using process data across the PISA cycle, describing, for each phase of the cycle, how process data can be leveraged to improve the quality of PISA (Section 2). The paper then highlights the challenges arising from the collection and use of process data in PISA, in particular related to their valid interpretations, their analysis, and associated ethical concerns (Section 3). Section 4 concludes.

Section 1: What are process data?

6. **Process data** capture information on students' **response processes** during a testing situation, i.e. on the thoughts, behaviours and feelings that underlie and drive how students respond to items (Ercikan and Pellegrino, 2017^[2]; Hubley and Zumbo, 2017^[3]). For instance, response processes might include the strategies and approaches that students choose to solve a problem, their motivation and engagement with the item, their stress level, or whether the interface is confusing them. The term 'process data' thus covers any data which provide evidence on these processes.

7. More often than not, response processes are not readily observable in the process data and need to be inferred (when it is possible to do so). For instance, students' problem solving strategies could be directly observed through verbal reports from the students themselves (e.g., the student explains to an interviewer the reasoning behind his actions), or could be inferred from their recorded sequences of actions on the computer.

8. There are multiple types of process data, and multiple ways to collect them. Each kind of data and collection method has its own advantages and disadvantages in terms of a) how readily observable responses processes are; and b) how scalable it is, which is a function of both implementation cost and of whether the data product is standardised. We present here a non-exhaustive list of process data that can be collected in assessment settings (summarised in Table 1 below):

- **Students' verbalisations of their response processes** (e.g., telling their thought processes and feelings to an interviewer while or after answering a test). Providing students are willing to tell an interviewer about their response processes, this data has the potential to provide direct information on a vast range of response processes, including strategies, emotions, and motivations. However, the data collected is not standardised and this collection method is costly to implement, so for the moment it cannot be easily scaled and used to make inferences at the population level. The verbalisation process also takes up time for students, and so might reduce the efficiency of a test.
- **Direct observations of students' behaviours**, including facial and vocal expressions and gestures (e.g., by an interviewer present in the room, or through a video recording). This data can inform on some response processes that are observable from behaviours, for instance strategies (e.g., gaming the system by trying rapidly all answers until finding the correct one), frustration or disengagement. However, many response processes would remain unobserved. In addition, such observations are complex to standardise and the implementation cost makes it more appropriate for small-scale studies.
- **Students' eye-movement pattern**. Eye-tracking tools capture where and for how long students look on the interface, enabling to retrace their eye movement patterns, as well as pupil dilation data. Eye-tracking has long been used in cognitive science to make inferences about human cognitive processes, and researchers in assessment have started to leverage its power to collect data on students' response processes. For instance, whether students are looking at relevant aspects of the assessment or not tells us something about their understanding of the item and was found to be associated with performance in a graphic literacy task (Langenfeld et al., 2019^[4]). The data that we get is standardised. However given the current costs associated with eye tracking machines and the logistics, this method is at the moment better more appropriate for small-scale studies.

- **Students' physiological responses** (e.g., heart rate, electro-dermal activity, blood volume pulse, skin temperature or face capture software). This data can inform on students' states when completing the assessment, such as their anxiety levels (for instance, heart rate variability was found to mediate the relation between specific math anxiety and problem-solving speed in an arithmetic task (Tang et al., 2021^[5])). Data from such methods are standardised, but not readily scalable due to the cost and logistics of using physiological measurement tools.
- **Students' brain activity measures** (e.g., through brain imaging techniques such as functional magnetic resonance imaging (fMRI), electroencephalography (EEG) and magnetoencephalography (MEG) or functional near-infrared spectroscopy (fNIRS), each having their advantages and drawbacks). Again long used in research in cognitive science, brain activity measures can inform on students' cognitive processes when solving a task (for instance, they can provide indications of students' cognitive load (Fishburn et al., 2014^[6]), and have been used to estimate cognitive load in a listening comprehension assessment (Aryadoust, Foo and Ng, 2021^[7])). These methods provide standardised data but at a cost preventing scaling up.
- **Students' digital traces of performance (response logs)**, such as *timing data* (e.g., time on task, time to first action, response time, inactive time); *action logs* (e.g., mouse clicks and mouse tracking information, keystrokes); or *intermediate solution states* (before submitting the final solution). These data can inform on students' use of the test's features (affordances), on their sequences of actions when solving the test, on how fast they submitted their answer, whether they were disengaged, their progress throughout the task. For instance, students' response time to an item is used to identify rapid guessing behaviours, reflecting students' disengagement with the tasks (Wise, 2017^[8]). These data are standardised and are easily collected within digital based assessment through computer logfiles. Currently, they provide the only scalable solution to collecting process data in large-scale assessments such as PISA.

Table 1: Types of process data

| Process data | Collection method | Observed or inferred response processes | Scalability potential | Appropriate use in the context of ILSAs |
|---|--|--|---|--|
| Students' verbalisations | Cognitive interviews and think aloud protocols | All thought processes and feelings are observable provided the students are able and willing to communicate about them | Low (unstandardised data and high cost) | Small scale validation studies during the item development phase |
| Direct observation of students' behaviours | Observation and reports by a test administrator or video studies | Some thought processes and feelings can be inferred | Low (unstandardised data and high cost) | Small scale validation studies during the item development phase |
| Students' eye-movement patterns | Eye-tracking measurement tools | Some thought processes and feelings can be inferred | Low (standardised data but high cost) | Small scale validation studies during the item development phase |
| Students' physiological responses | Physiological measurement tools | Some thought processes and feelings can be inferred | Low (standardised data but high cost) | Small scale validation studies during the item development phase |
| Students' brain activity measures | Brain imaging tools | Some thought processes and feelings can be inferred | Low (standardised data but high cost) | Small scale validation studies during the item development phase |
| Students' digital traces of performance (response logs) | Recording and extraction from computer logfiles | Some thought processes and feelings can be inferred | High (standardised data and low cost) | All stages of the cycle |

9. The characteristics of each type of process data and associated collection method influence at which phase of the assessment cycle they can be used. Methods which are costly, logistically heavy or provide non-standardised data will not be appropriate to use in large-scale assessments beyond the early stage of item development (when running small-scale validation studies). On the other hand, the collection of response logs data is the only one which will be relevant in all the other stages of the assessment cycle. The next section expands on this and details how different types of process data could be used at each stage of the PISA cycle.

Section 2: Using process data across the PISA cycle

10. From defining test domains to reporting and disseminating the data, process data have the potential to considerably enhance the validity and value of PISA. This section reviews applications of process data to each stage of PISA, covering the definition of test domain in the framework, item development and scoring, test construction, data adjudication and quality checks, scaling and conditioning, and dissemination for secondary research. This review does not intend to be exhaustive, but rather to discuss, through specific examples, how process data could be leveraged to improve the quality of PISA. Figure 1 summarises the different uses identified in this review for each stage of the PISA cycle.

Figure 1. The uses of process data across the PISA cycle



Framework development

11. The uses of process data, as well as the methods we use to analyse them, depend on the purpose of the assessment, that is the claims the assessment is designed for. Not all assessments need to use process data: this depends on the nature of the assessed construct. The frameworks should define whether the assessment aims to make claims on processes, and whether these claims require using process data or can be made by simply evaluating the final outcome of students' work. The relationship between process and outcome is in fact often blurry, and in some cases we can use a correct response as a proxy for well-executed reasoning. Before designing items that use process data, the assessment designers have to evaluate how central processes are to the interpretation of student performance, and how data on processes will be used in reporting.

12. Bergner and von Davier (2018^[9]) proposed a level structure that is useful to describe the different ways process data can be used, on a spectrum that goes from peripheral to central to the assessment claims:

- **Level 1: The process is irrelevant or at least ignorable given the outcome.** For example, if the purpose of measurement is to evaluate whether students can perform a one-step mathematical operation (e.g. the sum of two numbers), then the observation of whether students can achieve the correct response is sufficient.
- **Level 2: The process is auxiliary to the outcome** and may be understood through independent measures. One common example of this use consists in measuring whether students demonstrated a sufficient level of engagement in the task, computing the time they spent on the item. In a reading or mathematics test, engagement is not the same thing as the assessed construct, but process-based measures of engagement can help understanding differences in performance.
- **Level 3: The process is essential to understanding the outcome.** At this level, the scores assigned to an item account for process features. For example, in an item that asks the student to conduct and analyse results from experiments in a virtual science laboratory, the process data can be used to assign different scores to students who arrive at a solution methodically conducting the required experiments, and to those who use haphazard trial and error.
- **Level 4: The process is the outcome, and process scores are derived from an expert rubric.** Here we want to make claims on the capacity of students to perform some specific processes, so capturing and interpreting process data is essential to the validity argument of the assessment. An example is the assessment of self-regulated learning, that is a central construct in the PISA 2025 assessment of Learning in the Digital World. Self-regulated learning is by definition a process that unfolds as people plan, implement, monitor and evaluate their study strategies: it follows that the direct observation through process data of how test-takers engage in those activities is essential to make defensible claims on their competence in this construct. The scoring rubrics indicate how to interpret sequences of actions: for example, a rubric may establish, based on theoretical assumptions, that a student who tests his solution to an engineering problem after each substantial modification of the design demonstrates proficiency in self-regulated learning.
- **Level 5. The process is the outcome, and process scores are derived from a measurement model that accounts for dependencies in sequential data.** Differently from level 4, here dynamic probabilistic models are used to deal with the interdependence of observed actions in an extended task. For example, in Mystery Plants, an inquiry task about the effect of sunlight and fertilizer on the growth of different plant types, students progress through three different phases (Lin, Olivera-Aguilar, and Jia, 2015). In the ‘predict’ phase, they are asked to correctly predict whether different plants need different amounts of sun-light (to assess their prior knowledge about the object of the investigation). In the ‘observation’ phase, they can conduct experiments within a simulation to observe how plants react to light. In the final ‘explain’ phase, they have to communicate their findings. Students’ performance in each phase is dependent on the previous phases, so the most appropriate way to derive measures involves the statistical modelling of these dependencies, for example using Bayesian networks.

13. Most of the uses of process data in PISA and in other large-scale assessments have been at level 2. For example, measures of reading fluency based on response time data have been constructed to interpret reading performance. The scarce use of process data at level

3 and above is mostly due to the success of the simpler testing models from the past: ignoring processes that occur during testing simplifies the design of the tasks and allows the use of measurement models that require independent observations. However, this limited use becomes problematic whenever the assessments intend to make claims on reasoning and problem solving processes, given that these processes can only be imperfectly measured using response data. To what extent do the PISA constructs, as described in the frameworks, necessitate the use of process data?

14. Consider first scientific literacy. The most recent version of the PISA Science Framework defines scientific literacy as ‘the ability to use scientific knowledge and information interactively’ ([EDU/PISA/GB\(2022\)8](#) – PISA Science framework draft), suggesting that students should not just know about science but also be able to ‘do science’. Doing science involves some hands-on interaction with the natural world; that is, the design of an appropriate experiment or field observation, the construction of models to represent phenomena and the use of these models to make predictions. People learn to think like scientists interactively, trying different actions and observing their effects and consequences: assessing where students are in the development of this literacy thus requires to observe, using process data, to what extent students are capable of performing inquiry practices to solve authentic problems.

15. Technology-based, multimedia environments offer opportunities to present students with complex, life-like situations in which they can pursue a sustained investigation. Observing learners’ choices within the course of a guided inquiry in these digital environments can provide insights into students’ thinking without the interruption of a series of questions and answers. Because tasks can be designed so that students engage in multiple phases of inquiry (for example, planning an investigation into the quality of the water in a given watershed; collecting water samples within a simulated environment; organising and analysing the data they have collected; forming conclusions and communicating their findings), we can tap not just the individual inquiry “abilities” as described in the PISA Science framework, but also students’ ability to orchestrate these abilities during a complex, real-life activity. For example, WestEd SimScientist simulations challenge learners to explore interactively how to balance the proportions of algae, shrimp, and alefish to keep the food chain and environment healthy (National Center for Education Statistics, 2012_[10]). The environment employs multiple linked representations, such as variables under learner control, a runnable ecosystem model, graphs depicting changes in these variables, and summary data tables from values of the running model. Students can be assessed in their design of experiments and controls, in their graph interpretations, in identifying functional relationships (such as sketching a food web in an ecosystem), and so on.

16. Processes of thinking and problem solving constitute a central component of mathematics literacy. The 2022 iteration of the PISA Mathematics framework puts an emphasis on mathematical reasoning as the overarching, target construct. Mathematics reasoning ‘involves evaluating situations, selecting strategies, drawing logical conclusions, developing and describing solutions, and recognising how those solutions can be applied’ ([EDU/PISA/GB\(2018\)19](#)).

17. Many environments used for mathematics assessments, including those used in PISA, provide limited means for students to express themselves mathematically (Drijvers, 2019_[11]). Equation editors, graphing tools, geometry construction tools, statistical tools are either not available or very basic and difficult to use. At the current state of technology use in large-scale mathematics assessments, students can do more in a paper-and-pen environment than in a computer-based one, because on paper they can sketch and write whatever they want. There are, however, ample opportunities for innovation. The digital

environments that have been developed for online learning (e.g. Geogebra) offer opportunities for the design of rich, dynamic and interactive mathematics items, where students can express themselves mathematically in appropriate and sophisticated ways. Think of items in which students can construct graphs and geometrical objects, explore properties of such objects, and change them. As in the science domain, these new design tools invite students to “do mathematics” during the test, producing work that is close to authentic mathematics practice in a professional context.

18. Technological progress is also helping to make sense of the complex process data from these environments in a cost-efficient way, replacing the need for human raters. For example, computer algebra systems can interpret numerical and algebraic expressions (Sangwin and Köcher, 2016_[12]), and boolean variables in dynamic geometry systems can automatically score geometrical constructions (Kovács, Recio and Vélez, 2018_[13]). The so-called domain reasoners identify the steps a student makes, and can determine not only if the step is correct, but also if it brings the student closer to the solution, or rather represents a detour. This combination of interactive environments for mathematics expression and automated scoring procedures has the potential of producing more insightful assessments of students’ mathematic reasoning skills, that assign more weight to the effectiveness of the problem-solving process rather than to the fluency in elementary procedures.

19. In PISA reading, revisions of the assessment frameworks have reflected the increasing frequency of online reading. When students read online, they engage in a self-directed process of construction of meaning in a networked information space, that involves choosing what to read and connecting multiple online texts (Kiili and Leu, 2019_[14]). One practice that differentiates online reading from reading on paper is the opportunity to conduct informal research of the source in order to uncover potential biases (a strategy called lateral reading). While it is in principle possible to use traditional, static items to assess whether students engage in lateral reading (for example, asking directly to students what they would do to verify a source), a more reliable assessment strategy is based on the continuous observation of choices students make in online environments (for example, using the process data to verify whether students do internet searches to discover more information on the source).

20. The importance of integrating process data as sources of evidence is also evident in the innovative domains. Consider, for example, the assessment of collaborative problem solving. Much of the work addressing collaborative performance has focused on evaluating the final performance of the whole team or has used multiple-choice questions after the end of the collaborative activity (Kirschner et al., 2011_[15]). These methods provide limited information on how individuals manage the complex interactions among group members during the task activity (von Davier and Halpin, 2013_[16]). An assessment of collaborative skills should indeed evaluate students’ capacity to engage in productive interaction patterns, by working jointly on the task, sharing ideas and resources, and engaging with their collaborators’ suggestions. The PISA test of collaborative problem solving has moved one step forward in the assessment of collaboration processes by simulating interactive work between the tested students and computer avatars. However, in PISA the interactions were limited to opportunities to choose what to say to the avatar among a set of options in a scripted dialogue. Environments where real students collaborate on the same task and affordances for communication are available on demand, for example through chatbots or sharable screens, are more likely to elicit valid information on how well students work in a team. The complex data arising from open collaborative environments are now easier to convert into evidence through the application of natural language processing methods.

21. These examples show that the PISA domains put a strong emphasis on the processes of reasoning, learning, making and solving problems. The recent updates to the constructs have strengthened the centrality of processes in the assessment claims. Methods of measurement in PISA have only limitedly evolved to take into account these changes in the framework, continuing to rely almost exclusively on measures of the correctness of responses. Such misalignment generates validity issues. An important step forward would be to map in each assessment framework the practices/skills that are best measured through process data, indicating what evidence we expect to gather from these data and how this evidence is used to make claims.

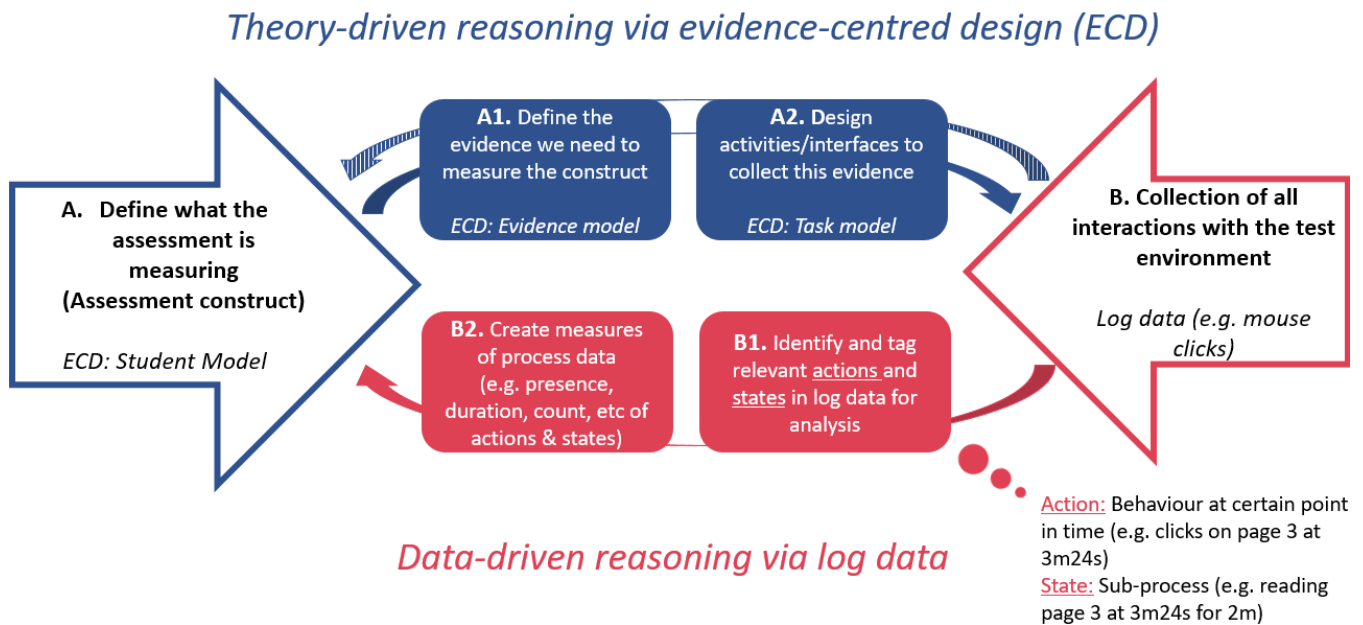
Item development and scoring

Using process data by design following the ECD framework

22. As previously described, using process data can present new opportunities for assessing competences that involve cognitive and behavioural processes. Once the claims about these processes are defined in the conceptual framework of the assessment, the next step is to develop substantive, technical and operational specifications of the assessment to gather the relevant evidence. The best approach for accomplishing this is to follow an evidence-centred design (ECD) framework (Mislevy, Almond and Lukas, 2003^[17]; Goldhammer et al., 2021^[18]). The ECD framework involves developing a student, evidence, and task model in a systematic way to produce accurate inferences about a student's competencies. For assessments that aim to make claims about process-related constructs, process data should be considered throughout the design phase. Process data "by design" ensures the evidence needed to make these claims is collected and the inferences made from this evidence are valid and reliable (Maddox, 2023^[19]).

23. A major advantage of computer-based assessment is that its software can capture a continuous stream of students' interactions with the digital interface. Each interaction is saved as a 'log' file. Log file data in and of itself is not evidence. However, when meaningful sequences of log data are identified, they can reveal important aspects of a student's work process, such as the presence (or not) of a solution strategy. Translating log file data into valid measures of a competence requires the data to be linked theoretically and empirically to the target assessment construct(s). Test developers can do this by: a) defining the different facets of a competence, how these facets are related and jointly contribute to student performance (student model); b) defining what evidence is required for making claims on each facet, developing scoring rules that convert observations into evidence in specific items, and then applying a statistical model to derive performance score(s) based on observations across all items (evidence model); and c) designing the task interfaces and activities to collect the observations specified in the evidence model (task model). This theory-driven construction of process indicators is described by the blue arrows in Figure 2: construct analysis provides the theoretical foundations to define what kind of evidence is needed and then to design activity spaces that allow to elicit the desired actions. In existing applications, this theory-based path to evidence identification is often complemented by data-driven construction of process indicators (see the red part of Figure 2), that typically relies on supervised or unsupervised data-mining (Goldhammer et al., 2021^[18]). Supervised methods identify those log-data sequences that can best predict relevant outcomes, such as success on the task. In unsupervised approaches, the logs obtained from an item are clustered to learn about underlying structures in the data. An individual's membership to a certain cluster is then interpreted in terms of a personal attribute of the test-taker (e.g. engaged or not engaged) or as an indication of a certain solution strategy.

Figure 2: Iterative design to use process data for assessment measurement



Adapted from Goldhammer et al. 2021

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

24. The PISA 2025 Learning in the Digital World (LDW) assessment aims to measure students' capacity in three components: 1) Computational and scientific inquiry practices; 2) Metacognitive monitoring and cognitive regulation processes; and 3) Non-cognitive regulation processes. This innovative assessment is constituted by items that are connected to the same scenario, as in other PISA assessments. Within each item, however, the provided evidence takes the form of not only the final work product, but also considers the test-taker's behaviour over time capturing attributes of the work process (e.g., presence or absence of a solution strategy).

25. In the unit 'I like that' developed for the LDW framework, students are asked to build part of a computational model by investigating and establishing relationships between variables (Figure 3). More specifically, students investigate the factors that determine how an individual rates a movie (favourably or poorly) using an experimentation tool. Based on their experiments, they use a conceptual mapping tool to build an underlying model that associate characteristics of movie to ratings. This model will be used by a digital application to recommend movies to users.

Figure 3: Example task from the PISA 2025 Learning in the Digital World prototype unit "I like that!"

I like that! Example

Complete the model.

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

Experiments

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

Add Experiment

Model Check work

Characteristics:

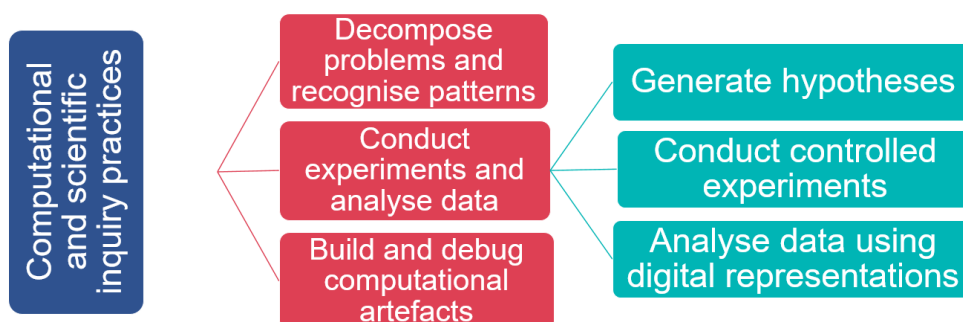
- Release Date
- Cinema Distance
- Friends' Reviews

Ticket Price (+) → Movie Rating

1. The student model

26. According to the ECD framework, the student model defines the variables (knowledge, skills, and other attributes) that are involved in target constructs and their relationships between the variables. Figure 4 shows an example of a student model for computational and scientific inquiry practices. The structure of the student model contains three first-level variables and several second-level variables for each first-level variable.

Figure 4: Example student model for 'Computational and scientific inquiry practices'



27. The test designers have developed a construct map for each variable of the student model. A construct map describes a student’s ability and understanding of the facet, including common misconceptions, at different levels along a progression (Wilson, 2009^[20]) According to this theoretical model, a student’s ability to ‘conduct experiments and analyse data’ can be determined, in part, by their ability to use the control of variables strategy (CVS). In “I like that!” (Figure 3), students must use the experimentation tool to

conduct experiments in which they use CVS. By running multiple experiments in which they vary the values of the target independent variable ('ticket price') and keep all other variables constant (in this case, 'distance of cinema'), students can draw valid inferences about the impact of the target independent variable on the dependent variable ('movie rating'). Once the student model variables are defined, the next step is defining the evidence that is needed to make these claims.

2. The evidence model

28. The evidence model consists of scoring rules and a statistical model (A1 in Figure 2). Scoring rules make explicit the operations to evaluate the quality of the work product (e.g., responses, sequence of actions). The statistical model specifies how the different scores are related and should be aggregated to make claims on the variables in the student model.

2a. Development of scoring rules

29. When using process data to generate evidence of process-related constructs, there are two main steps to follow: 1) identifying the actions (behaviour at a certain point in time) and states (transitions between events, or 'sub-processes') that are meaningful to describe and interpret the work process, and 2) combining these actions and/or states into process indicators of the work process. Process indicators can be generalised across different task types and unit scenarios. For each process indicator, there must be rules to stipulate how it will be computed. These rules are generally defined by expert rubrics that indicate what score to assign to specific sequences of actions (e.g.: when students do x and then y , then assign a score of z). These scores can be treated as isolated observations (that can be used for scaling together with response data) or connected into more complex measurement models that specify relationship between the different observations and account for dependencies in sequential data (e.g., probabilistic process models using Bayes nets; (Bergner and von Davier, 2018_[9])).

30. What does this look like in practice? For Step 1, the assessment software can be programmed to identify meaningful actions and states and to record them in a specific format. For actions, the format generally includes a timing convention (e.g., millisecond the action is observed), the type and location of the action (e.g., change value for variable in row 2, run a test) the result of the action (e.g., variable changed from 1 to 2). In the LDW example task, here are two example lines from the log file of a student's continuous stream of log events:

```
"startTime": "DATE 14:01:45.115"}, {"data": {"event_prev": null, "event_type": "dropdown_change", "event_result": 1, "event_location": "distanceofcinemaRow2" }
```

```
"startTime": "DATE 14:01:46.795"}, {"data": {"event_prev": null, "event_type": "dropdown_change", "event_result": 2, "event_location": "ticketpriceRow2" }
```

31. Using knowledge about the LDW example task (Figure 3) and the assessment system, the data can be translated into time-stamped actions:

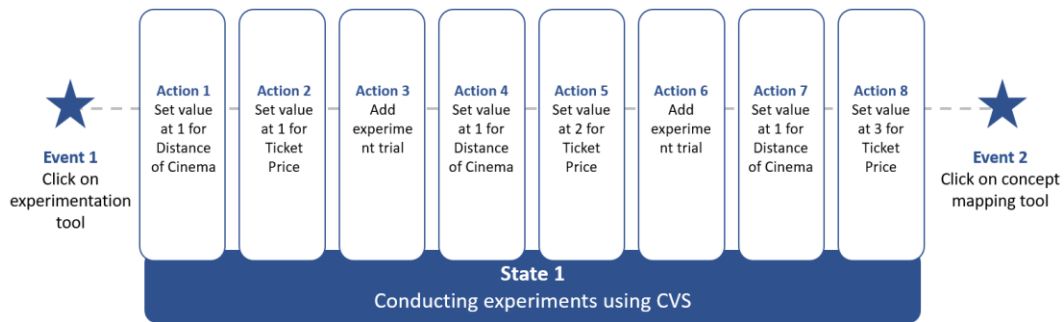
- Action 1: At time 14:01:45.115, 'in row 2 of the experiment table, change the empty value of the first variable (Distance of cinema) to a value of 1'
- Action 2: At time 14:01:46.795 'in row 2 of the experiment table, change the empty value of the second variable (Ticket price) to a value of 2'

32. States identify meaningful events that occur either at the initiative of the student or the system itself. When an event that has been deemed meaningful occurs, the software is

programmed to record a snapshot of the workspace. The comparison of students’ work between two states can provide evidence of the existence of a solution strategy.

33. For Step 2, sequences of actions and/or states are then aggregated to become evidence of a process-related variable (i.e., ‘process indicators’). In the LDW example, a process indicator for conducting controlled experiments can be based on the observation of whether students have followed an expert-defined sequence of actions using the experiment tool. As shown in Figure 5, in this case we store all the ‘actions’ that the student does on the experiment tool, and verify whether the student has applied the CVS strategy.

Figure 5: Process of carrying out control of variable strategy to find a relationship between two variables



34. Partial credit rules can be developed to award recognition to students whose process data reveal they have understood the logic of controlled experiments but who made some procedural mistake (for example, testing less than 3 values for the independent variable). Table 2 provides an example evidence rule table for how to transform log-file data into a scoring rule of CVS for the target variable.

Table 2: Example evidence rule table for conducting controlled experiments

| Attribute of work process | Actions and/or states (can be task-specific) | Process indicator (generalisable across units) | Question-based scoring rules for “I like that” unit (task-specific) |
|--|---|--|--|
| Use control-of-variables strategy | Actions: <ul style="list-style-type: none"> Change value Add experiment State: <ul style="list-style-type: none"> Conducting experiments A, B, C... for X duration of time | Student conducts a controlled experiment by running multiple trials that meet the following conditions: <ul style="list-style-type: none"> For the independent variable, a range of values are selected across trials For the control variable, the same value is selected across trials | Does the student conduct three experiments in which they vary ticket price while keeping the values of distance of cinema constant? No: 0 Yes: 1 |

35. Pilot studies should be conducted to validate and refine the evidence model and associated scoring rules. Data-driven approaches, such as data-mining techniques, can analyse log file data to search for appropriate solution processes that were not already accounted for. If new process indicators can be derived from these techniques, they would then need to be mapped to the assessment construct using theory or domain-expert knowledge (see striped, blue arrow in Figure 2; (Mislevy et al., 2012_[21]; Goldhammer et al., 2021_[18]))

36. In LDW, there are several tasks with more open solution spaces than the controlled experimentation task we just described. Data-driven exploration could refine and capture

the solution paths that may not have been considered in the theoretical definition phase of the evidence rules.

37. At this stage in development, it is important to consider the threat of construct-irrelevant variance. Unexpected variation in student behaviour can weaken the validity of the inferences made from process indicators about student proficiency in the target construct (Goldhammer et al., 2021_[18]). In other words, there may be other variables at play that could explain differences in the observed behaviours of a given process indicator. In the LDW task example (Figure 3), one instance of construct-irrelevant variance could be the inability of the student to conduct CVS (or any experiment at all) because they are unable to use the drop-down menus of the experimentation tool. Validation studies, including cognitive laboratories, usability studies, and quantitative pilot studies, can be used to identify such sources of construct-irrelevant variance. There are several ways to reduce these types of threats, such as further defining the process indicator or adjusting the task following feedback from the validation studies.

2.b Development of a statistical model

38. The second component of the evidence model is the statistical model that relates the scoring rules to the variables. It does this by combining scores across tasks to derive a final score. The choice of statistical model should be based on theoretical consideration and adapted to the nature of the data. From a theoretical perspective, domain experts can define expectations for how a low, medium and high achiever would score on each task, how the indicators map to each of the latent variable and to the overall, higher-level construct. This theory might indicate that some indicators are more or less important to the assessment of student ability in a particular facet, and help assign weights to each indicator.


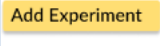

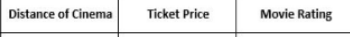

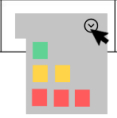


39. The model must account for dependencies of measures across tasks because, in an assessment like LDW, the items are not fully independent. Students progress from easier to more difficult tasks and they are expected to ‘learn’ concepts and tools. These dependencies can be addressed by using complex analytical models, such as tree-based item response models (e.g., IRTrees) and Bayesian networks.

3. The task model

40. The task model should specify the objectives, prompts and interface features of an assessment activity, consider and explain design choices that impact task difficulty, and describe the optimal solution(s) that demonstrate proficiency (Mislevy et al., 1999_[22]). Each task must be designed to elicit the evidence for making inferences about the assessment construct, as defined in the evidence model. To document these decisions, multiple models of tasks are generally developed for an assessment, in order to ensure a variation of interfaces and activities and thus strengthen the claims made about students’ abilities.

41. In the LDW example (Figure 3), the interface (see row B in Table 3) includes an experimentation tool with clickable elements, designed to be as accessible as possible. In this task model, the difficulty can be manipulated by varying the number of variables, the number of values a variable can take, the type of functional relationship between the input and output variables, the presence of moderating variables (input variables that influence the relationship between another input variable and the outcome).

Table 3: Partial task model for conducting controlled experiments

| Task model elements | Task design and considerations | Relevant visual reference |
|--|--|---|
| A. Task prompt | The task is to “conduct experiments to find out how ticket price impacts movie rating” | |
| B. Interface features | Experimentation tool designed as a table that automatically ‘runs’ when values are chosen |  |
| | ‘Add experiment’ button for students to conduct multiple experiment trials |  |
| C. Design choices that impact task difficulty | Number and complexity of relationships to investigate |  |
| | Number of control variables to account for |  |
| | Presence (or not) of moderating variables | |
| | Number of values students need to keep constant/vary |  |
| | Choice of values (closed choice via dropdowns, or free input like typing) |  |
| | Variables and order of variables present in the experimentation tool (pre-filled with relevant variables in appropriate columns, or allowing students to choose/change variables themselves) |  |
| D. Optimal solution to first task | When the input variables can take three values, student completes three experiment trials where they vary the value of the input variable (e.g. ticket price) keeping constant values of the other input variables (e.g. distance to cinema) |  |

Providing additional diagnostic information about students

42. Beyond allowing better measures of process-based constructs, process data can provide additional diagnostic information about students. This information might not be used for the main performance scale, but reported separately in the form of additional indicators that help interpret a student’s or a system’s position in the scale. Data-mining methods can uncover useful information about how student groups differ in their solution strategies, in the way they manage their effort on the tasks (e.g., productively persisting, wheel-spinning or giving up), or in the types of mistakes they do (Greiff, Wüstenberg and Avvisati, 2015^[23]; Salles, Dos Santos and Kespaik, 2020^[24]). Identifying and describing these differences can help policymakers and teachers find concrete ways to adjust their support. For instance, performing cluster analysis on students’ process data from an interactive mathematics item from a large-scale assessment in France, Salles, Dos Santos and Kespaik (2020^[24]) found that students fell in different types of profiles based on the solution strategies they adopted (trial and error, operational approach to the concept of function, versus a targeted, structural approach to functions). While these strategies could not explain students’ achievement, such results could provide valuable information to teachers and curriculum designers.

Test construction

43. Beyond the definition of test domains and supporting the improvement of items during the development phase, process data from log-files have the potential to support and refine test construction. This covers the selection of items that will make it to the final item pool, how these final items are assembled into test forms and blocks, and the inclusion of accessibility features into the test.

Informing the selection of the final item pool

44. Based on item statistics gathered through the Field Trial, items are selected for the Main Survey. This selection currently involves checking item functioning, for instance inspecting correlations between item and total scores, item-by-country interactions, and coder reliability (OECD, 2020_[25]). Technology-based assessments considerably enlarge the array of information on item characteristics which can be collected and which could be used to refine the selection of items for the final item pool in PISA.

45. For example, at the test construction stage, test developers use assumptions about the time students would need to answer different types of items in order to ensure that students have enough time to complete the assessment. Here, timing data at the item-level can be leveraged to check and validate these assumptions, and refine item selection. For instance, Hicks, Circi & Sikali (2021_[26]) analysed data from the mathematics assessment in NAEP to compare students' actual median response time to multiple choice questions (MCQ) to the conventional assumption used by test developers that MCQ are completed in 60 seconds. While their findings confirms that most MCQs could be responded within 60 seconds (independently from the level of difficulty), their analysis identified outlier items for which median response time was much higher than a minute.

46. As another example, log files enable to track students' response changes to items, such as switching from option A to B in a MCQ before validating option B. This information can cast light on distractors which elicit an excessive amount of changes, thus complementing the traditional distractor analysis which solely looks at how often wrong answers are selected and validated by students. Response change data can be leveraged to further explore items which exhibit differential item functioning when selecting items for the final item pool (Circi, 2021_[27]).

Informing test assembly

47. Once the final item pool has been selected, items are assembled into test forms, or blocks. Test assembly denotes the process of pulling together different items from the pool to constitute test forms or blocks, controlling the order in which the items are presented (item position) and the test duration in order to ensure that the test is balanced and not speeded. The process data collected through technology-based assessments provide additional information on each item, which could be used to inform and refine test assembly.

48. First, data on students' level of (dis)engagement could be used in test assembly, for instance to better distribute items across test forms and blocks, or to provide insights on how we could improve the test design to reduce disengagement. Analysing PISA 2018 data, Avvisati et al. [EDU/PISA/GB(2023)5] have studied the item characteristics which are associated with higher rates of rapid guessing (indicating disengagement). Their study shows that response format is one of the strongest predictors of rapid guessing, with simple multiple-choice items being more than 20 times more likely to trigger rapid guessing behaviour compared to open-response items, and more than 13 times more likely than complex multiple-choice items. The inclusion of certain types of multimedia as well as

content length, was also associated with rapid guessing. Besides, their results demonstrate a strong effect of the break between the first and second testing hour on rapid guessing, such that the level of engagement at the beginning of the first and second testing sessions is much higher compared to those items administered just before the break. These results thus highlight the potential gains of reducing the length of the testing sessions either by decreasing the total number of items, or by increasing the number of breaks during the session.

49. Second, in the context of multi-stage adaptive testing (MSAT), which PISA has introduced for the major domains starting from 2018, process data could help improve the assembly of test blocks. For instance, using simulations on PISA 2018 data, Chang and Yi have shown that incorporating response time in an on-the-fly MSAT (OMST) has the potential to improve measurement efficiency (Chang and Yi, forthcoming^[28]) (study carried out as part of the RDI MSAT project).

Enhancing accessibility

50. Lastly, process data could inform test construction by providing insights on the uptake and use of accessibility features available in the testing platform (such as text-to-speech or highlighting text) and accessibility options made available to students (such as extended time accommodation) – which is difficult to observe and record with paper-and-pencil tests. The analysis of accessibility features and accommodations usage can enable us to know whether students actually use them, and whether their use is associated with improved performance. Such results can then inform the test construction stage on whether incorporating these features in the assessment is beneficial or not, and could provide guidance on the optimal way of integrating such affordances.

51. For instance, Lee, Hicks, & Circi (2020^[29]) looked at the use of text-to-speech feature in the NAEP mathematics test, to understand which students use it when it is made available to all students according to universal design framework. They found that 38% of accommodated students (who would have traditionally had access to read aloud accommodation) used it at least once. In addition, 44% of non-accommodated students used text-to-speech at least once. Using propensity score matching, where accommodated students who used text-to-speech were matched with accommodated students who did not use it (and similarly with non-accommodated students), they found that the performance of accommodated students who used text to speech was higher than those who did not use it. However, the reverse was found for non-accommodated students: the performance of non-accommodated students who used text-to-speech was slightly lower than those who did not use it. This might indicate that text-to-speech may be a distraction for students who do not need it, or who are not used or prepared to using it, and thus would further research before making it available to all students in PISA.

Data adjudication and quality check

52. After data collection from the PISA Main Survey, each national dataset is reviewed as part of the data adjudication process. This review is mainly based on meeting the PISA sampling standards, the outcomes of the translation process, the PISA Quality monitoring visits, the quality and completeness of the submitted data, which includes concerns about the data quality identified during scaling and in preparation for reporting (PISA 2018 Technical Report, Chapter 14). By providing information on response patterns, time and location, process data could be judiciously integrated in this procedure to evaluate the data collection integrity and serve as an important decision-making element in the data adjudication process.

Identifying data fabrication

53. First, process data can be used to identify instances of data fabrication. For instance, abnormally low response times in a given country, or response times which are inconsistent with the mean times in other countries, can be used to detect data falsification. Looking at data from PIAAC, Yamamoto and Lennon (2018_[30]) have identified in one country an unusual pattern where high proficiency was associated with abnormally low response times on the literacy items, which raises concerns about the possibility that these responses were falsified.

54. In addition to response time, the precise timing of responses can be extracted and used as an additional element to spot data falsification. Thus, Yamamoto and Lennon (2018_[30]) identified a country case in PIAAC where many duplicate responses were found (scattered across a wide range of participants), and each of these duplicate responses happened to have the same precise timing – to the millisecond. Such a pattern is highly unlikely to arise in the context of the PIAAC adaptive design, which raises concerns regarding the likely fabrication and data collection integrity in this country.

Detecting dysfunctions during test administration

55. Moreover, in the current transition to online testing, process data can also help detect dysfunctions which happened during testing, such as instability of internet connection. For instance, response time and the number of visits to single items, can be used to indicate anomalies at the item level. Thus, an issue leading to restarting the session might lead to a negative unit duration. Similarly, the sum of duration of each item can be checked to see if it is inferior to the unit duration (which should be the case); as well as the score change in relation to the number of visits to an item (van Rijn, 2020_[31]).

Development of proficiency measures

56. In PISA, group-level proficiency distributions are estimated with a latent regression model composed of i) a measurement model which uses IRT to estimate how students' performance at the test depends on proficiency (scaling); and ii) a population model which uses a latent regression to estimate how proficiency relates to students' background information (conditioning). Including process data at this stage could improve the accuracy of group-level proficiency measures.

Improving the measurement model

57. With the advent of technology-based assessments, the research on the development of IRT models using response time has developed. However, this did not translate yet on actual test design for existing or new assessments.

58. Using PISA 2012 mathematics test data, Reis Costa and colleagues (2021_[32]) investigated the utility of integrating time-on-task data along with response accuracy in a joint framework. Their results show that using response time in a joint model significantly improved measurement precision compared to using the standard (response accuracy only) model, and this was true for all countries. However, they found that time-on-task parameters may differ across countries, both in terms of how much time students spent on items overall, and in terms of how much time students spent on specific items. In addition, while their results show that using a joint model did not substantially affect the overall assessment of proficiency levels for the different countries, the estimated correlations between the individual ability estimates varied across models, thus suggesting that the joint model does not capture exactly the same abilities as the standard one. This suggests that

additional validation research is needed to assess how to integrate timing information in the measurement model.

59. Several researchers have approached this question from a disengagement perspective. For instance, using students' response time, Wise and DeMars (2006_[33]) developed an IRT model which incorporates student's effort as measured by rapid guessing behaviour. They propose that students facing a testing situation engage either in a solution behaviour or a rapid-guessing behaviour. This engenders different item response functions (one for the solution behaviour, and one for the rapid-guessing), which is not taken into account in classic IRT models. Combining the two item response functions into a single model moderated by response strategy (the "effort-moderated" IRT model), they show that their approach can provide more accurate item parameter estimates and proficiency estimates with greater validity than with a standard IRT model, even when the proportion of rapid guessing is low (e.g., 2%).

60. Another way to take into account response time data in the scaling model is by informing the coding of omitted responses for the purpose of the IRT modelling. Students might omit responses on a test for several reasons, such as low skills, low motivation, or lack of time. How these omitted responses are coded (incorrect or missing) influences the outcomes of the IRT model. In PISA, these are coded as incorrect. However, omitted responses coded as incorrect can induce a negative bias into estimates of item difficulty as well as a negative bias in estimates of group means (Rose, von Davier and Xu, 2010_[34]). At the same time, omitted responses coded as missing (assumed missing at random and excluded from the estimation) induce larger standard errors. This is where response time could be used to differentiate between rapid omitted responses which are unrelated to students' proficiency and could be treated as missing, and omitted responses which are related to proficiency and could be treated as incorrect in the IRT scaling (Weeks, von Davier and Yamamoto, 2016_[35]).

Improving the population model

61. Process data can also be included as additional predictor variables in the latent regression (population) model. Since 2018, PISA incorporates response time (as person-level deciles) in the population modelling to generate plausible values. This contributes to improving the measurement precision of proficiency (Annex K and H of PISA 2018 Technical report (OECD, 2020_[25])).

62. Beyond response time, behavioural indicators, actions, sequences and other types of timing data could also be used in population models. However, due to the massive amount of such data, and since population models already handle a very large number of background variables, adding information from process data needs to go hand in hand with an improvement of the variable selection process (von Davier et al., 2019_[37]). Additional research is needed to understand how process data could best be summarised and incorporated in population models.

Dissemination

Enriching the reports of PISA results

63. As developed in the section 2.1 (item development), the collection of process data in PISA has the potential to make PISA results much more relevant to the research community and education stakeholders, for instance by highlighting students' strengths and weaknesses and providing fine-grained diagnostic information on students' mistakes and strategies to solving an item. For instance, the PISA 2012 international report contained

a case-study with analyses of students' reading fluency, their persistence, and their navigation behaviour based on the analysis of process data from a single unit of the computer-based reading assessment. Among other results, these analyses showed that students from some countries are significantly more likely to read more slowly, and showed that navigation patterns was related to success in the task (OECD, 2015^[36]). Similarly, the PISA report "21st Century Readers: Developing Literacy Skills in a Digital World" used process data to analyse students' navigation patterns in a reading unit. Based on the number of pages students visited, their navigational behaviour and use of hyperlinks, and the time they spent on the different pages, the report identified four different groups of students corresponding to different navigation types. These types of navigation were found to be related with performance in the task (reading achievement) (OECD, 2021^[37]).

Facilitating secondary research exploring further questions

64. Beyond the production of reports by the OECD, the vast amount of process data collected in PISA could tremendously benefit the research community if made available to researchers for carrying secondary analysis. Such secondary analyses can explore research topics related to students' learning that have not been addressed in the PISA reports. For instance, (Greiff, Wüstenberg and Avvisati, 2015^[23]) carried secondary analyses of logfile data from the complex problem solving assessment in PISA 2012 to study the relationship between the vary-one-thing-at-at-time (VOTAT) strategy and students' performance and to identify groups of students with different levels of non-mastery. Another such example is the study carried by Teig, Scherer and Kjærnsli (2020^[38]), which used Norwegian log-file data from the PISA 2015 science test, focusing on two scientific inquiry tasks. Their study revealed three distinct profiles of students' inquiry performance, which were associated with different exploration behaviour, inquiry strategy, time-on-task, and item accuracy. In addition, these three profiles showed different demographic characteristics (gender, socio-economic status, and language at home), attitudes (enjoyment in science, self-efficacy, and test anxiety), and science achievement.

65. Besides complementing international reports, secondary analyses of process data also have the potential of testing and trialling improvements that could enhance PISA methods for the future cycles. For instance, (Chang and Yi, forthcoming^[28]) used PISA 2018 data, including response time data, to carry a simulation study which assessed the effect of incorporating response time in an on-the-fly multistage adaptive testing design using PISA 2018 data. The study concludes that such a design is promising.

66. Of course, it is crucial that the process data are disseminated in a way that facilitates research. Indeed, there are many ways that actions and states can be formulated in the logging system and aggregated into process indicators. This can create barriers to comparing data across research projects and for researchers from different domains to analyse the data. Defining a standardised way to log these features and remaining transparent about how process indicators are created can increase the use of the data for secondary research, and thus raise the overall value of the assessment. This implies a clear documentation on the actions and states for which log-data are collected, on how the different process indicators are constructed. Using a standardised labelling and organisation of the dataset can also facilitate secondary research.

Section 3: Challenges to the use of process data in PISA

67. While the use of process data could be of tremendous value in the future developments of PISA, it presents some key challenges that need to be considered and addressed. These concern, in particular, difficulties regarding the validation of process data interpretation, challenges around the ethical aspects of collecting, using and disseminating process data, as well as complexities with the statistical modelling of dependent observations.

Validity challenges

68. As mentioned in Section 1, response processes are most of the time not readily observable in the process data and need to be inferred. Making such inferences from process data variables and indicators is not straightforward and needs to be defensible and validated with theoretical arguments and empirical evidence, in the same way that test score interpretation needs to be validated (AERA, APA and NCME, 2014^[39]).

69. Process data variables and indicators can indeed be open to multiple interpretations and reflect different behaviours or thought processes. For instance, process data indicating single actions could reflect different behaviours depending on the time and context in which they occur. Rapid guessing at the beginning of a test might indicate low engagement, but it might also indicate test speededness when occurring at the end of a test (Ercikan, Guo and Por, forthcoming^[40]). Therefore, multiple types of process data should be crossed (for example, using retrospective verbalisations to validate an interpretation of a log-data sequence) and additional sources of information might need to be gathered to help interpret process data variables and indicators.

70. Besides, behavioural patterns can vary across different groups of students, which adds another layer of complexity to the interpretation of process data. Indeed, sequences and speed of actions might depend on individual characteristics such as students' ability, experience, culture or language. For example, using data from the PISA 2018 Science test, Guo and Ercikan (2020^[41]) have shown that the thresholds for flagging rapid responses differ across nine linguistic and cultural groups – which is assumed to be partly due to differences in the efficiency of the input editor method used to input text in the different languages. This suggests that using a single threshold to identify rapid responses, as is usually the case, inaccurately categorises responses of students from certain groups of countries as rapid response (or the reverse). This points to the importance of taking into account students' characteristics when analysing and interpreting process data.

71. Clear standards and guidelines on how process data should be used are currently lacking, since the existing standards references (such as the AERA, APA and NCME's Standards for Educational and Psychological Testing) preceded the rise of process data. Thus, the use of process data in large scale assessments still “lacks the explicit professional ‘framing’ and warrant that is provided for more conventional scores from assessment through various sets of standards” (Murchan and Siddiq, 2021^[42]).

Ethical challenges

72. The collection of a multitude of process data poses new ethical challenges for large scale assessments. In their comprehensive review study, Murchan and Siddiq (2021^[43]) note that the enthusiasm that has accompanied the rise of process data has so far directed most of the efforts on improving the data collection and analysis techniques, at the expense of designing and incorporating ethical safeguards for participants. These ethical challenges include concerns related to privacy and data protection – which have concentrated most

of the efforts on ethics of process data so far –, but also issues regarding transparency and consent, as well as the minimisation of adverse impacts for participants.

73. First, the collection of process data raises concerns regarding data protection and privacy. Indeed, some of the process data collected could potentially be used to identify students. For instance, the use eye-tracking technology presents risks of privacy loss as gaze patterns can be used to infer sensitive information such as physical or mental health status, drug consumption or sexual orientation, and even to uniquely identify individuals (Kröger, Lutz and Müller, 2020_[44]). The processing of data from recordings and video studies similarly present obvious privacy risks which need to be mitigated. Even data from logfiles such as keystrokes can be used to identify a participants' identity, based on their typing manner and rhythm – an identification method known as keystrokes dynamics or typing biometrics (Lu et al., 2020_[45]).

74. Second, because process data have emerged as a by-product of the switch from paper- to technology-based assessments, they have so far mostly been collected unbeknownst to participants. This raises concerns relative to transparency and fairness of the tests, especially if the process data are used to create performance scores or diagnostic indicators. Indeed, it is important that all stakeholders are provided with information about what is being recorded, how they are being assessed and how the collected data will be used.

75. Lastly, it is important to assess if the use of process data might create adverse impacts for students. For instance, using process data in the scoring process without accounting for potential differences in which different groups of students may respond to assessment tasks could threaten the assessment's fairness (as well as validity) by disadvantaging these groups. In particular, students' experience with digital technology needs to be taken into account.

Analysis challenges

76. Lastly, the use of process data presents challenges for data analysis, first of which is the particular structure and lack of standardisation of these data compared to data from traditional assessments. Indeed, in open and interactive environments, students can follow different sequences of actions. One action determines whether other actions are possible or not. For example, the observation on whether a student corrected his response following a hint depends on the student's action of asking for a hint. As a consequence, items cannot be assumed to be independent, since one captured event will result from previous actions made by the student. In addition, this creates non-random missingness in the data. The challenge is therefore to develop appropriate measurement models which accurately represent the data, accounting for non-random missing data and local dependencies between items. Examples of such models which deserve further research are tree-based item response models, or IRTrees (Boeck and Partchev, 2012_[46]; Jeon and De Boeck, 2015_[47]), which takes into account the covariance within complex tasks through node structures that capture sequential processes.

77. Second, we have seen that including timing data in the measurement (IRT) model has the potential to improve the accuracy of item parameter and proficiency estimates. At the same time, doing so may complexify and alter the definition and interpretation of proficiency, as Reis Costa and colleagues (2021_[32]) have shown.

78. In addition, the most accurate way to incorporate response time into IRT model, and especially, its assumed relationship with response accuracy, is far from clear. Indeed, several IRT models assume that response time and accuracy are dependent processes (van der Linden, 2016_[48]); however different models contrast on the nature of this relationship

(is greater accuracy associated with increased, or decreased response times? Does this relationship depend on ability level?). In a study, Embretson (2021^[49]) found that the correlations between participants' response times with variables such as mean item response time, item difficulty, test position, relative item difficulty and item cognitive complexity vary widely across participants, thus suggesting that selecting a single response time model that would fit a variety of students for a given assessment is far from straightforward. Under these uncertainties, incorporating timing data in scaling models may compromise the measurement of trends and the comparability of computer-based PISA (which uses the process data) with paper-based PISA (without process data). Previous research has shown that changes in the scaling approach can indeed affect comparability across time and countries: for example, looking at data from the PISA mathematics tests from the four rounds between 2003 and 2012, Heine and Robitzsch (2022^[50]), have found that different analytical decisions for scaling that were made in the various cycles had a small to decisive influence on country ranking and trend estimates.

79. Finally, we have suggested that process data could also be leveraged to augment the population model and serve as additional predictors of proficiency in the latent regression. However, as underlined in Section 2, the vast amount of process data that is collected in logfiles poses a considerable challenge in doing so. Since population models already incorporate numerous background variables, augmenting them with process data increases the risk of possible overparameterisation. Therefore, additional research on variable selection processes is needed to identify how process data could best be incorporated in population models (von Davier et al., 2019^[51]).

Section 4: Conclusions

80. This paper has shown that using information on the responding process can serve different assessment purposes in the context of PISA. Borrowing information from log-data through the use of an evidence-centred design process, we could better measure complex constructs such as scientific inquiry, lateral reading or collaboration skills, that are defined by processes of work and not only by end results. Log-data can also be used to provide diagnostic evidence on students' strengths and weaknesses in the target competencies, informing on the type of strategies they used to solve a problem or the type of mistake they made. Information from log-data is also essential for detecting and responding to test disengagement. Furthermore, process data can be leveraged to improve test construction by providing additional data on item characteristics which can be used to enhance test assembly; and they could support the improvement of accessibility by providing information on the uptake and impact of affordances embedded in the testing platform. Besides, in the context of the current transition to online delivery of the PISA test, process data can be used to assess the integrity of the data collection in order to identify data falsification or fabrication, as well as technical anomalies during administration that might affect results. Finally, integrating timing data into scaling models have the potential to improve measurement precision.

81. However, these advantages come hand in hand with important challenges. This paper has identified three main difficulties associated with the rise of process data in large scale assessments. First is a validity challenge, as process data can reflect various behaviours and their uses and interpretations thus need to be validated theoretically and empirically. Second is an ethical challenge, as the use of process data raises important concerns on data protection, transparency and consent. We should also be careful not to introduce potential adverse impacts for students. Third is an analysis challenge related to the complexities of modelling dependent process data observations and to the consequences of the integration of process data in scaling and population models on measurement quality and comparability across countries and time.

82. Taking full advantage of process data in PISA therefore requires to adopt a systematic strategy to address these challenges, covering the whole assessment cycle. To start with, it would be very important to carefully assess whether process data can improve the measurement of target constructs or generate useful diagnostic information. If so, item design should follow a 'process data by design' approach based on Evidence Centred Design (ECD). As part of this process, it would be crucial to adopt sound validation strategies – correlational and experimental (Goldhammer et al., 2021^[18]) – and to conduct multiple rounds of validation studies to validate score interpretation, spot potential issues with the task, and improve item content and design. Uncertainties regarding the proper use of process data in scaling and population models call for additional research and reflection on the advantages and disadvantages of incorporating them. Routines and criteria for identifying technical issues during test administration and data fabrication with process data should also be developed. Furthermore, in order to facilitate secondary research it would be important to that process data files are easily understandable and usable by external researchers. This implies the production of clear documentation on which process data are collected and how process indicators are constructed from the raw data, as well as clear and consistent labelling of process variables. Finally, addressing the ethical challenges demands that students are informed about what data is being recorded and on what they are being assessed; and that process data which could be used to identify students or which could reveal sensitive information are not released as part as public datasets.

References

- AERA, APA and NCME (2014), *Standards for Educational and Psychological Testing, 2014 Edition*. [39]
- Aryadoust, V., S. Foo and L. Ng (2021), “What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments?”, *Language Testing*, Vol. 39/1, pp. 56-89, <https://doi.org/10.1177/02655322211026876>. [7]
- Bergner, Y. and A. von Davier (2018), “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>. [9]
- Boeck, P. and I. Partchev (2012), “IRTrees: Tree-Based Item Response Models of the GLMM Family”, *Journal of Statistical Software*, Vol. 48/Code Snippet 1, <https://doi.org/10.18637/jss.v048.c01>. [46]
- Chang, H. and Z. Yi (forthcoming), “On-the-fly Multistage Testing: An Alternative Design for PISA”. [28]
- Circi, R. (2021), *Examination of response change in Multiple Choice Items*. [27]
- Drijvers, P. (2019), “Digital assessment of mathematics: Opportunities, issues and criteria”, *Mesure et évaluation en éducation*, Vol. 41/1, pp. 41-66, <https://doi.org/10.7202/1055896ar>. [11]
- Embretson, S. (2021), “Response Time Relationships Within Examinees: Implications for Item Response Time Models”, in *Springer Proceedings in Mathematics & Statistics, Quantitative Psychology*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-74772-5_5. [49]
- Ercikan, K., H. Guo and H. Por (forthcoming), “Use of Process Data in Advancing Practice and Science of Technology-Rich Assessments”, in *Innovative assessments*, OECD Publishing. [40]
- Ercikan, K. and J. Pellegrino (2017), *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*, Routledge. [2]
- Fishburn, F. et al. (2014), “Sensitivity of fNIRS to cognitive state and load”, *Frontiers in Human Neuroscience*, Vol. 8, <https://doi.org/10.3389/fnhum.2014.00076>. [6]
- 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>. [18]
- Greiff, S., S. Wüstenberg and F. Avvisati (2015), “Computer-generated log-file analyses as a window into students’ minds? A showcase study based on the PISA 2012 assessment of problem solving”, *Computers & Education*, Vol. 91, pp. 92-105, <https://doi.org/10.1016/j.compedu.2015.10.018>. [23]
- Guo, H. and K. Ercikan (2020), “Differential rapid responding across language and cultural groups”, *Educational Research and Evaluation*, Vol. 26/5-6, <https://doi.org/10.1080/13803611.2021.1963941>. [41]

- Heine, J. and A. Robitzsch (2022), “Evaluating the effects of analytical decisions in large-scale assessments: analyzing PISA mathematics 2003-2012”, *Large-scale Assessments in Education*, Vol. 10/1, <https://doi.org/10.1186/s40536-022-00129-5>. [50]
- Hicks, J., R. Circi and E. Sikali (2021), “Where data meet assumptions: Visualization of multiple-choice item response time”, *Educational Measurement: Issues and Practice*, Vol. 40/1, p. 6. [26]
- Hubley, A. and B. Zumbo (2017), “Response Processes in the Context of Validity: Setting the Stage”, https://doi.org/10.1007/978-3-319-56129-5_1. [3]
- 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>. [47]
- Kiili, C. and D. Leu (2019), “Exploring the collaborative synthesis of information during online reading”, *Computers in Human Behavior*, Vol. 95, pp. 146-157, <https://doi.org/10.1016/j.chb.2019.01.033>. [14]
- Kirschner, F. et al. (2011), “Differential effects of problem-solving demands on individual and collaborative learning outcomes”, *Learning and Instruction*, Vol. 21/4, pp. 587-599, <https://doi.org/10.1016/j.learninstruc.2011.01.001>. [15]
- Kovács, Z., T. Recio and M. Vélez (2018), “Using automated reasoning tools in GeoGebra in the teaching and learning of proving in geometry”, *International Journal for Technology in Mathematics Education*, Vol. 25/2, https://doi.org/10.1564/tme_v25.2.03. [13]
- Kröger, J., O. Lutz and F. Müller (2020), “What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking”, in *Privacy and Identity Management. Data for Better Living: AI and Privacy, IFIP Advances in Information and Communication Technology*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-42504-3_15. [44]
- Langenfeld, T. et al. (2019), “Integrating Multiple Sources of Validity Evidence for an Assessment-Based Cognitive Model”, *Journal of Educational Measurement*, Vol. 57/2, pp. 159-184, <https://doi.org/10.1111/jedm.12245>. [4]
- Lee, S., J. Hicks and R. Circi (2020), *Insights on Text-to-Speech as a Universal Design Feature: NAEP Mathematics Process Data*. [29]
- Lu, X. et al. (2020), “Continuous authentication by free-text keystroke based on CNN and RNN”, *Computers & Security*, Vol. 96, p. 101861, <https://doi.org/10.1016/j.cose.2020.101861>. [45]
- Maddox, B. (2023), “The uses of process data in large-scale educational assessments”, *OECD Education Working Papers*, No. 286, OECD Publishing, Paris, <https://doi.org/10.1787/5d9009ff-en>. [19]
- Mislevy, R., R. Almond and J. Lukas (2003), “A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN”, *ETS Research Report Series*, Vol. 2003/1, pp. i-29, <https://doi.org/10.1002/j.2333-8504.2003.tb01908.x>. [17]

- 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. [21]
- Mislevy, R. et al. (1999), “A cognitive task analysis with implications for designing simulation-based performance assessment”, *Computers in Human Behavior*, Vol. 15/3-4, pp. 335-374, [https://doi.org/10.1016/s0747-5632\(99\)00027-8](https://doi.org/10.1016/s0747-5632(99)00027-8). [22]
- Murchan, D. and F. Siddiq (2021), “A call to action: a systematic review of ethical and regulatory issues in using process data in educational assessment”, *Large-Scale Assessments in Education*, Vol. 9/1, <https://doi.org/10.1186/s40536-021-00115-3>. [43]
- Murchan, D. and F. Siddiq (2021), “A call to action: a systematic review of ethical and regulatory issues in using process data in educational assessment”, *Large-scale Assessments in Education*, Vol. 9/1, <https://doi.org/10.1186/s40536-021-00115-3>. [42]
- National Center for Education Statistics (2012), *NAEP: Looking Ahead—Leading Assessments into the Future*. [10]
- OECD (2021), *21st-Century Readers: Developing Literacy Skills in a Digital World*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/a83d84cb-en>. [37]
- OECD (2020), *PISA 2018 Technical Report*, OECD Publishing, Paris. [25]
- OECD (2015), *Students, Computers and Learning: Making the Connection*, PISA, OECD Publishing, Paris, <https://doi.org/10.1787/9789264239555-en>. [36]
- Provasnik, S. (2021), “Process data, the new frontier for assessment development: rich new soil or a quixotic quest?”, *Large-scale Assessments in Education*, Vol. 9/1, <https://doi.org/10.1186/s40536-020-00092-z>. [1]
- Reis Costa, D. et al. (2021), “Improving the Precision of Ability Estimates Using Time-On-Task Variables: Insights From the PISA 2012 Computer-Based Assessment of Mathematics”, *Frontiers in Psychology*, Vol. 12, <https://doi.org/10.3389/fpsyg.2021.579128>. [32]
- Rose, N., M. von Davier and X. Xu (2010), *Modeling nonignorable missing data with item response theory (IRT)*, Res Rep ETS RR-10-11. Educational Testing Service. [34]
- Salles, F., R. Dos Santos and S. Keskaik (2020), “When didactics meet data science: process data analysis in large-scale mathematics assessment in France”, *Large-scale Assessments in Education*, Vol. 8/1, <https://doi.org/10.1186/s40536-020-00085-y>. [24]
- Sangwin, C. and N. Köcher (2016), “Automation of mathematics examinations”, *Computers & Education*, Vol. 94, pp. 215-227, <https://doi.org/10.1016/j.compedu.2015.11.014>. [12]
- Tang, J. et al. (2021), “Respiratory Sinus Arrhythmia Mediates the Relation Between “Specific Math Anxiety” and Arithmetic Speed”, *Frontiers in Psychology*, Vol. 12, <https://doi.org/10.3389/fpsyg.2021.615601>. [5]
- Teig, N., R. Scherer and M. Kjærnsli (2020), “Identifying patterns of students’ performance on simulated inquiry tasks using <scp>PISA</scp> 2015 log-file data”, *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1400-1429, <https://doi.org/10.1002/tea.21657>. [38]

- van der Linden, W. (2016), “Lognormal response-time model”, in van der Linden, W. (ed.), *Handbook of item response theory*: Taylor & Francis Inc. [48]
- van Rijn, P. (2020), *Uses and Reporting of Process Data*. [31]
- von Davier, A. and P. Halpin (2013), “COLLABORATIVE PROBLEM SOLVING AND THE ASSESSMENT OF COGNITIVE SKILLS: PSYCHOMETRIC CONSIDERATIONS”, *ETS Research Report Series*, Vol. 2013/2, pp. i-36, <https://doi.org/10.1002/j.2333-8504.2013.tb02348.x>. [16]
- von Davier, M. et al. (2019), “Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments: An Overview of Challenges and Opportunities”, *Journal of Educational and Behavioral Statistics*, Vol. 44/6, pp. 671-705, <https://doi.org/10.3102/1076998619881789>. [51]
- Weeks, J., M. von Davier and K. Yamamoto (2016), “Using response time data to inform the coding of omitted responses.”, *Psychological Test and Assessment Modeling*, Vol. 58, pp. 671-701. [35]
- 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>. [20]
- Wise, S. (2017), “Rapid-Guessing Behavior: Its Identification, Interpretation, and Implications”, *Educational Measurement: Issues and Practice*, Vol. 36/4, <https://doi.org/10.1111/emip.12165>. [8]
- Wise, S. and C. DeMars (2006), “An Application of Item Response Time: The Effort-Moderated IRT Model”, *Journal of Educational Measurement*, Vol. 43/1, pp. 19-38, <https://doi.org/10.1111/j.1745-3984.2006.00002.x>. [33]
- Yamamoto, K. and M. Lennon (2018), “Understanding and detecting data fabrication in large-scale assessments”, *Quality Assurance in Education*, Vol. 26/2, <https://doi.org/10.1108/QAE-07-2017-0038>. [30]