Received: 10 March 2022 | Revised: 19 June 2023 | Accepted: 21 June 2023

DOI: 10.1111/jcal.12847

ARTICLE

Journal of Computer Assisted Learning WILEY

20 years of interactive tasks in large-scale assessments: Process data as a way towards sustainable change?

Matthias Stadler[1,2] | Laura Brandl[1] | Samuel Greiff[3]

[1]Department of Psychology, Ludwig Maximilians University Munich, Munich, Germany

[2]Institut für Didaktik und Ausbildungsforschung in der Medizin, LMU Klinikum, Ludwig-Maximilians-Universität München, Munich, Germany

[3]Department of Behavioral and Cognitive Sciences, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Correspondence

Matthias Stadler, Klinikum der Ludwig-Maximilians-Universität München, Institut für Didaktik und Ausbildungsforschung in der Medizin, Pettenkoferstr 8a, München 80336, Germany.

Email: matthias.stadler@med.uni-muenchen.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: COSIMA; DFG-Forschungsgruppe 2385

Abstract

Background: Over the last 20 years, educational large-scale assessments have undergone dramatic changes moving away from simple paper-pencil assessments to innovative, technology-based assessments. This comprehensive switch has led to some rather technical improvements such as identifying early guessing or improving standardization.

Objectives: At the same time, process data on student interaction with items has been shown to carry value for obtaining, reporting, and interpreting additional results on student skills in international comparisons. In fact, on the basis of innovative simulated assessment environments, news about student rankings, under- and overperforming countries, and novel ideas on how to improve educational systems are prominently featured in the media. At the same time, few of these efforts have been used in a sustainable way to create new knowledge (i.e., on a scientific level), to improve learning and instruction (i.e., on a practical level), and to provide actionable advice to political stakeholders (i.e., on a policy level).

Methods: This paper will adopt a meta-perspective and discuss recent and current developments with a focus on these three perspectives. There will be a particular emphasis on new assessment environments that have been recently employed in large-scale assessments.

Results and Conclusions: Most findings remain very task specific. We propose a necessary steps that need to be taken in order to yield sustainable change from analysing process data on all three levels.

Implications: New technologies might be capable of contributing to the research-policy-practitioner gap when it comes to utilizing the results from large-scale assessments to increase the quality of education around the globe but this will require a more systematic approach towards researching them.

KEYWORDS

large-scale assessments, process data, replication, research-policy-practitioner gap, technology-based education

1 | INTRODUCTION

Over the last 20 years, educational large-scale assessments have undergone dramatic changes moving away from simple paper-pencil assessments to innovative, technology-based assessments (von Davier et al., 2019). One example of the emergence of technology-based large-scale assessments is the Programme for International Student Assessment (PISA). This program, arguably the most extensive international educational assessment program, started to partially collect their data through technology-based assessment of science literacy back in 2006 (OECD, 2010). In 2009 PISA already administered another core competence using technology (digital reading assessment; OECD, 2012). This was extended in 2012 for the third core competence (mathematical literacy) plus adding a technology-based problem-solving assessment (OECD, 2013). In 2015, technology-based assessment was the primary mode of assessment in PISA (OECD, 2017). One reason was the inability to design authentic, interactive, and dynamic tasks for 21st-century skills with traditional paper-pencil formats (OECD, 2010). Technology-based assessments make the use of multimedia, simulations, interactive tasks, and virtual reality possible (Goldhammer et al., 2020). In addition to allowing for the operationalization of previously unobtainable competencies, using technology-based assessments allows for continuous measurement of the response process (i.e., process data), instead of only discrete states of responses depicted through the answers given to a task (i.e., product data; Thille et al., 2014).

This comprehensive switch from paper-pencil assessments to technology-based assessments has led to some rather technical improvements such as identifying early guessing (e.g., Kong et al., 2007) or improving standardization of assessment and scoring (e.g., Goldhammer et al., 2020). At the same time, process data on student interaction with items have been shown to carry value for obtaining, reporting, and interpreting additional results on student skills in international comparisons (e.g., Reis Costa et al., 2021; Xiao et al., 2021). Process data was used to relate behaviour to cognitive processes (Greiff et al., 2016), to validate score interpretations (Kane & Mislevy, 2017), and led to a better theoretical understanding of the construct under investigation (Goldhammer et al., 2017; Goldhammer & Zehner, 2017).

However, few of these efforts have been used sustainably to decrease global inequalities, and realize universal quality education (United Nations Educational, Scientific and Cultural Organization [UNESCO], 2015) by creating new knowledge, improving learning and instruction, and providing actionable advice to political stakeholders (Dawson et al., 2019). This paper will adopt a meta-perspective and discuss recent and current developments focusing on these three perspectives. There will be a particular emphasis on new assessment environments that have been recently employed in large-scale assessments and how they might contribute to the research-policy-practitioner gap when it comes to utilizing the results from large-scale assessments to increase the quality of education around the globe.

2 | INTERACTIVE TASKS IN LARGE-SCALE ASSESSMENT

2.1 | New types of assessment

One of the driving forces behind the fast and comprehensive switch from paper-pencil assessments to technology-based assessments in international large-scale assessments has been the need to assess so-called 21st-century skills (Care et al., 2012). These 21st-century skills encompass a set of skills deemed critically important to student success in today's world, particularly as students move on to college, the workforce, and adult life, such as solving complex problems individually and collaboratively or possessing the media literacy to utilize and critically evaluate digital sources of information. Assessing these competencies requires assessment tools that respond to the test-taskers' inputs to allow for adequately complex and realistic tasks. Unlike conventional tasks (such as multiple-choice questions), these interactive tasks change, while the test-taker is trying to solve them, providing feedback to interventions or new information (Stadler et al., 2015). For instance, collaborative problem-solving tasks would hardly be valid if there was no interaction between the test-takers and the collaboration partners (Stadler, Herborn, et al., 2020). Likewise, assessments of hyper-text reading (reading and understanding digital text organized in a non-linear hypertext format) need to allow the test-takers to choose what information they want to read actively and in what order (Hahnel et al., 2023).

All of these new forms of assessment share that the interaction between test-takers and the assessment are expressed in observable actions (e.g., mouse clicks, eye-movements, keyboard inserts). Researchers are, thus, no longer limited to measuring the final outcome of an assessment (i.e., product data) but can also investigate the steps and actions resulting in the specific outcome through analyzes of test-taking behaviours (i.e., process data; Greiff et al., 2016; He et al., 2021).

2.2 | Process data and sequence data

In contrast to product data, process data and sequence data is seen as empirical information depicting behaviour that leads to the measured outcome (Goldhammer & Zehner, 2017). Typical process data are response times or the number of actions taken, whereas sequence data, as a special form of process data, describes the qualitative action sequences that lead to a specific result (Pohl et al., 2021; von Davier et al., 2019). Sequence data hence includes timing data, adding a quantitative dimension. Analysing process data and sequence data instead of only product data allows insights into the process leading to the eventual outcome. Researchers have already used process data to answer research questions as diverse as the detection of early guessing behaviour (e.g., Kong et al., 2007), validation of product data (e.g., Kane & Mislevy, 2017), early identification of students at risk to show inadequate performance (e.g., Wolff et al., 2013), analyses of incorrect responses and reasons (e.g., Ulitzsch et al., 2021) and a better theoretical understanding of the construct under investigation

(Goldhammer et al., 2017; Goldhammer & Zehner, 2017). Accordingly Pohl et al. (2021) argue that test-taking behaviour is not a nuisance factor that may confound measurement, but an aspect that provides important information on how examinees approach tasks, which is relevant for real-life outcomes.

Regarding the use of sequence data, Greiff et al. (2018) reported that students might show similar overall performance and yet can be distinguished according to their strategic behaviours in the tasks. These results indicate that process indicators depict individual differences in the ability that are not necessarily depicted in product data. This interpretation was further corroborated on laboratory data by Stadler, Hofer, and Greiff (2020), who found that participants solving a set of complex problem-solving tasks systematically differed in both time-on-task and number of clicks despite having reached the same outcome. This difference in behaviour was systematic and represented differences in ability as indicated by significant relations to an external criterion (participants' GPA). Moreover, the differences in behaviour could be explained by adjusting their effects on participants' GPA for individual differences in general problem-solving ability, which reduced them to negligible levels. He and von Davier (2016) used sequence data from the Programme for the International Assessment of Adult Competencies (PIAAC) studying how action sequences from problem-solving tasks are related to task performance finding several distinct action sequences that were related to correct responses (such as actions related to using software-tools).

While there is a surge of interest among researchers in harnessing process data, this rich resource's full utilization through dedicated analyses remains in its embryonic stages (Stadler et al., 2019). Significant improvements have been made in employing process data to enhance scoring accuracy and reporting in educational large-scale assessments (Pohl et al., 2021). However, the real value of integrating interactive tasks into these programs lies in their unique ability to capture action sequences that facilitate an exploration of the underlying reasons for students' success and failure (von Davier et al., 2019).

These interactive tasks provide a distinctive opportunity to contribute to the development of more sophisticated models of student cognition. By yielding detailed sequence data, we gain a more granular understanding of how students approach and navigate through different tasks. This allows us to observe the evolution of their problem-solving strategies over time, providing empirical evidence that can validate or challenge existing cognitive theories. Such insights can then directly inform the design of more nuanced, targeted instructional methods and learning materials, thus enriching the teaching-learning process.

Despite the evident value, comprehensive analyses employing this resource are scarce, often restricted to single or a few selected items with little common theoretical underpinning and minimal attempts at replicating findings. In the second part of this paper, we will discuss how these missed opportunities have resulted in a lack of sustainable change in education at the scientific, practical, and political levels. As we move forward, it is essential to shift our focus from merely improving scoring and extending reporting, towards fully exploiting the potential of interactive tasks in generating refined cognitive models that can transform educational practices and theories.

3 | ISSUES LIMITING SUSTAINABLE CHANGE

3.1 | Scientific level

A substantial obstacle preventing sustainable change, brought about by the use of interactive tasks in large-scale assessments at the scientific level, is the strong task-specificity of findings. Replications, already a rarity in educational research (Makel & Plucker, 2014), are virtually non-existent when it comes to sequence data from interactive tasks (c.f., Brooks et al., 2015 for a positive example). Several reasons may account for this, such as the relative infancy of the field. However, we contend that the lack of generalizability of findings and a missing relation between data and theory strongly limit the replicability of research on sequence data from interactive tasks in educational large-scale assessments, and thus, its scientific value.

Interactive tasks are often highly complex, involving multiple interrelated variables, usually embedded in a certain semantic context. These contexts only permit specific interactions between them and the test-takers. Thus, directly relating specific interactions with one item to interactions with other items becomes a challenging task, especially if these items do not even allow for these particular interactions. Many studies interested in comparing processes across items are therefore forced to rely on relatively low-level metric analyses (Ihantola et al., 2015), such as relating time-on-task or the number of interactions to the latent construct being assessed (e.g., Greiff et al., 2016).

Drawing on Mislevy's (2019) view, we suggest an explicit differentiation between low-level features and higher-level features. Low-level features, such as time-on-task or the number of interactions, can be more idiosyncratic and may not convey the same meaning across different tasks (Stadler, Radkowitsch, et al., 2020). On the other hand, higher-level features, derived from low-level ones, can present robust evidence that is pertinent across different tasks, thereby providing the possibility of conceptual replication even when items differ between studies.

An inventive solution to this predicament was offered by He et al. (2021), who related the performance on several PIAAC tasks to the distance of the observed behaviour sequence from an ex-ante defined ideal sequence. This approach allows for generalizing findings across various tasks that do not need to be similar as long as it is possible to determine an ideal sequence of actions. However, these findings would still exist within a theoretical vacuum as long as the ideal sequence is not linked to a theory-based definition of the construct.

As an application of the approach of distinguishing low-level and higher-level features, (Brandl et al., 2021) coded the interactions between learners and a training simulation for medical diagnoses based on theoretically defined diagnostic activities (Fischer et al., 2014). This focus on higher-level features allowed the study to move beyond task-specificity, thereby enabling the generalization of findings across various diagnostic tasks. Aggregating the process data in this way allows to train machine-learning algorithms to predict successful diagnoses in various diagnostic tasks. This study makes it apparent how relating process data to established theoretical concepts can make the findings generalizable. The diagnostic activities used to code the interactions are not specific to any individual task, and the same method could be applied to any diagnostic training simulation regardless of context. Lotz et al. (2017) demonstrated how intelligence relates to individual differences in interaction frequency and quality changes in a computer-based problem-solving task. The authors find that, while all test-takers improve their test-taking behaviour across tasks on average, individual differences in intelligence predict the speed and range of this improvement. This example illustrates how rather basic process information can still be related to theoretically defined latent constructs.

In conclusion, to truly advance scientific knowledge through process data analyses, it's essential to generalize findings from specific items and link them to established constructs (see also Kroehne & Goldhammer, 2018). This requires the capacity for replication, systematic testing of theories, and the understanding of the difference between low-level and higher-level features. Recognizing this differentiation sets a crucial theoretical consideration for the level of abstraction in analysing process data, paving the way for more substantial scientific progress. Additionally, as underscored by (Goldhammer et al., 2021), there is a pressing need to validate the interpretation of measures based on process data. Even when a theoretical link between data and theory is postulated, this link necessitates substantiation with theoretical and empirical arguments to ensure its validity. Therefore, the integration of theoretical considerations, the differentiation of low-level and higher-level features, and the validation of interpretations collectively form the pillars of more rigorous and impactful research in this domain.

3.2 | Practical level

Despite the scientific challenges described above, there are many high-quality studies on the use of process data in educational large-scale assessments, demonstrating the benefit of modelling new data sources and incorporating process data in the statistical modelling of multiple possible assessment data (He & von Davier, 2016; Jiang et al., 2021; Pohl et al., 2021; von Davier et al., 2019). Process data can help validate and facilitate measuring response accuracy and provide supplementary information in understanding test-takers' behaviours, the reasons for missing data, and links with motivation studies.

However, with the evolution of educational large-scale assessment from a paper-based technology to an electronic one, the focus of these assessments has evolved, too (Bennett, 2015). Over the past several decades, the most common use of educational assessment has been for institutional purposes such as state school accountability. Accordingly, lots of research on the use of process data has concentrated on this use of assessment. However, in recent years, the value of assessment as a feedback tool informing individual learning (formative assessment) has been realized (e.g., Chudowsky & Pellegrino, 2003; van der Kleij et al., 2015). Whereas testing to serve institutional purposes may not diminish in absolute terms, there is reason to believe it will diminish in relative terms as assessment to serve individual learning purposes becomes more frequent. The increasing prominence of formative assessment is being driven by many factors, including advances in measurement and data science and the emergence of electronic learning environments.

Obviously, international large-scale assessments are primarily designed to facilitate group-based assessments and comparisons across large populations, not individuals (von Davier et al., 2019). Nonetheless, they can and should support learning (Chudowsky & Pellegrino, 2003). Especially interactive tasks inherently offer a type of feedback to the test-takers through the evolution of the task in response to their interactions (Greiff et al., 2016). This feedback, as we conceive it here, does not correspond to traditional, evaluative feedback, but instead refers to the changing state of the task according to the decisions made by the test-takers. Essentially, the task environment responds and adapts based on the actions of the test-takers, thus providing them with an implicit form of feedback about the consequences of their actions within the task scenario. This results in learning opportunities that can be used more or less efficiently, which needs to be considered when using these tasks as a means of standardized testing. Rather than trying to reduce these learning opportunities by limiting the tasks responsiveness, it may be beneficial to assess individual learning rather than the mere ability to solve the task. To benefit individually from an assessment situation, especially from a complex interactive task, learners require individualized scaffolding (e.g., Azevedo et al., 2004). For example, educators could use process-data from computer-based assessments to differentiate specific behaviours in students and use this information to provide individualized support (e.g., Li et al., 2020). Accordingly, process data analyses have long been considered a promising tool to detect a need for scaffolding and provide individualized support, yet most of this potential remains essentially untapped in day-to-day teaching practice (Bakharia et al., 2016). Most previous studies have drawn on historical data to identify patterns in students' process data and related these patterns to academic performance, retention, or other institutional outcomes. Utilizing process data for individual learning purposes requires understanding the pedagogical context that influences student activities and how identifying patterns in students' learning behaviours can help influence and contribute to more positive learning experiences (Gašević et al., 2016; Lockyer et al., 2013). An essential next step in advancing the practical relevance of new assessment technologies in educational large-scale assessments will, therefore, be to align the design of assessment with learning design to use assessments not only for institutional information but also as a source of individual learning.

3.3 | Policy level

Finally, modern educational large-scale assessments are an increasingly important part of the educational research and policy landscape internationally (Rutkowski et al., 2013). For instance, PISA claims to have become “the world's premier yardstick for evaluating the quality,

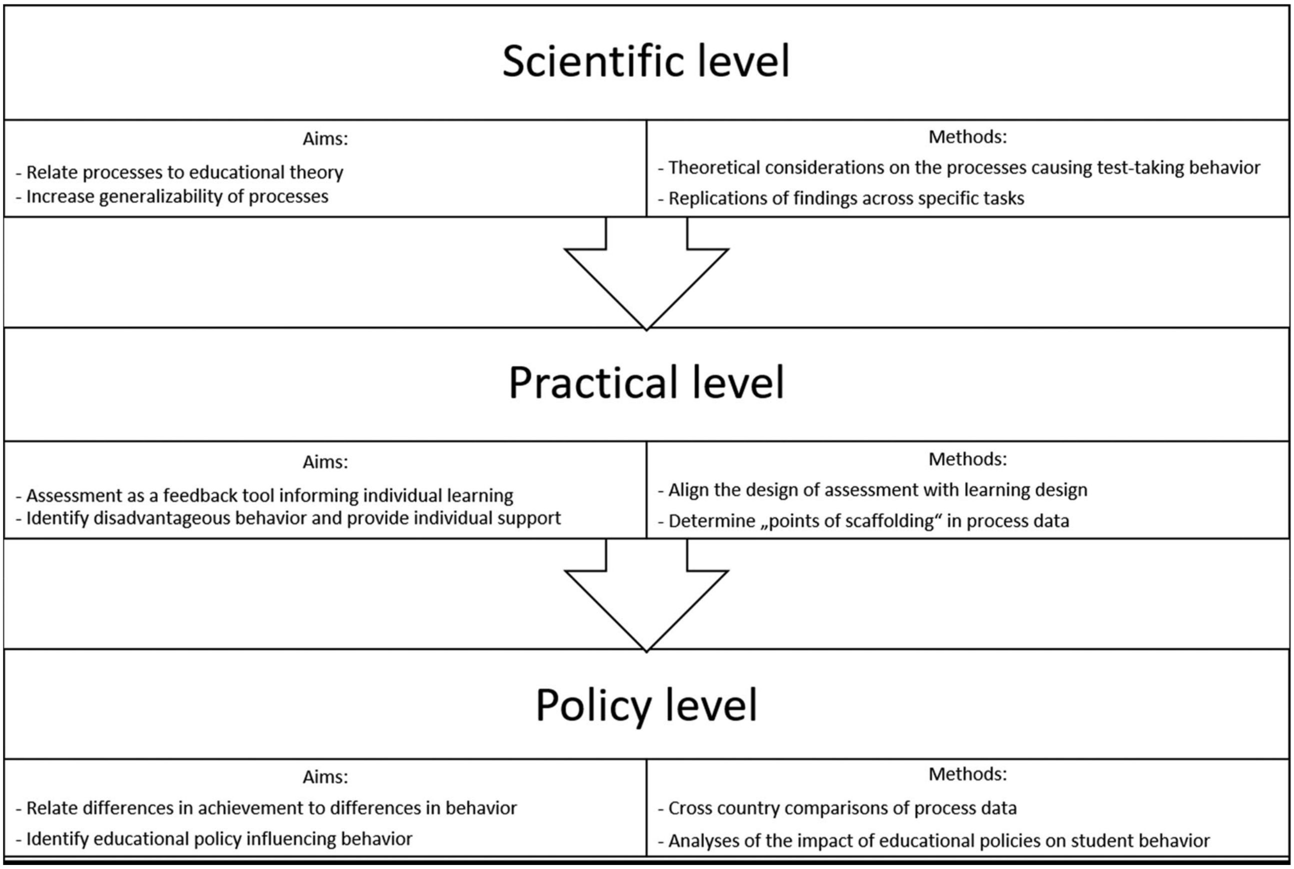
Scientific level
Aims:
- Relate processes to educational theory
- Increase generalizability of processes
Methods:
- Theoretical considerations on the processes causing test-taking behavior
- Replications of findings across specific tasks
Practical level
Aims:
- Assessment as a feedback tool informing individual learning
- Identify disadvantageous behavior and provide individual support
Methods:
- Align the design of assessment with learning design
- Determine „points of scaffolding" in process data
Policy level
Aims:
- Relate differences in achievement to differences in behavior
- Identify educational policy influencing behavior
Methods:
- Cross country comparisons of process data
- Analyses of the impact of educational policies on student behavior

FIGURE 1 The use of process data in educational large-scale assessment for sustainable change.

equity and efficiency of school systems" (OECD, 2016, p. 2). In fact, despite many criticisms and potential issues with comparability across countries (see e.g., Winthrop & Simons, 2013) as well as methodological constraints (e.g., Rutkowski & Rutkowski, 2016), there are several examples of how results from educational large-scale assessments have been converted into educational policy such as the formation, expansion and improvements to national assessment and evaluation systems, the revision of curriculum standards, often to include and emphasize PISA-like competencies, or promoting equity through school financing (Breakspear, 2012; Wagemaker, 2014).

However, in international large-scale assessments, the focus lies mainly on achievement test scores as a measure of competence, which makes sense when discussing performance on an aggregated level for quality monitoring (Skedsmo & Huber, 2017). Unfortunately, the idea that the criterion of competence is what someone can do often downplays the importance of how the person arrives at this competence (Havnes & Prøitz, 2016; Oliveri & Davier, 2014). In other words, merely relating aggregated sum scores to differences in educational systems such as the number of all-day schools or integrative schools without a very good understanding of the behaviour on which the test values are based seems problematic (e.g., Gür et al., 2012; Kuhlmann & Tillmann, 2009) and is unlikely to yield sustainable changes (Pohl et al., 2021).

Analyses of process data can help provide this understanding. For instance, Greiff et al. (2016), analysed log-file data of a complex problem-solving tasks for students from 44 countries and economies. The authors find that there were different levels of non-mastery that ranged from applying no systematic strategic behaviour to actually applying the appropriate strategy but still failing to solve the task. On the backdrop of these results, they discuss implications and future potentials of log-file analyses in educational large-scale assessments for researchers, teachers, and policy makers. This study demonstrates how for policy makers, interesting comparisons between educational systems might emerge from the relation between actual behaviour and overall proficiency. In the PISA 2012 cycle, for instance, Polish students performed reasonably well in mathematics (518 points in the international comparison of the PISA scale), science (526 points), and reading (518 points) but performed considerably worse in complex problem-solving compared with other countries or economies (481 points). Log-file analyses revealed how this performance drop could be explained by (a lack of) specific actions, for instance, because Polish students never learned the principle of isolated variation or because they were too reluctant to explore a problem situation comprehensively. This provides interesting starting points for policy decisions and educational priorities (see Greiff et al., 2015).

Questions such as what exactly it is that students do better in one country compared with another may provide insights into how teaching practices foster or neglect certain behaviours but need to consider cultural differences in learning and teaching (Huang et al., 2016).

4 | CONCLUSION

In summary, we posit that educational large-scale assessments, particularly through their evolution from simple paper-pencil tests to innovative technology-based and simulated environments, hold tremendous potential to advance research, educational practice, and policy-making. Despite this potential, the sustainable utilization of the rich information these assessments provide has been hindered, impacting their potential to enhance global education quality. As illustrated in Figure 1, there are necessary changes to be undertaken at the scientific level in how we analyse process data to foster sustainable changes at the practical and policy levels. Primarily, linking process data to educational theory is crucial for enhancing the generalizability of our findings. This link not only enables the utilization of assessment results as individual learning feedback tools but also allows the identification of disadvantageous behaviours, paving the way for targeted individual support.

To achieve this, the alignment of assessment design with learning theories is paramount. The process data emanating from such theoretically grounded and practically meaningful assessments can then elucidate achievement disparities across countries or educational systems. Policy predicated on such robust data can have a lasting, sustainable impact on students' education. This exploration underscores a fundamental need for further research dedicated to the sustainable, theory-driven utilization of process data from interactive tasks in large-scale assessments. Our aspiration is that this research will lead to systemic changes that bridge the gap between research, practice, and policy in education, ultimately contributing to the quality of education worldwide.

ACKNOWLEDGEMENTS

The contributions by Matthias Stadler and Laura Brandl was supported by a grant from the Deutsche Forschungsgesellschaft DFG (COSIMA; DFG-Forschungsgruppe 2385). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/jcal.12847.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study

ORCID

Matthias Stadler https://orcid.org/0000-0001-8241-8723

REFERENCES

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology*, *29*(3), 344–370. https://doi.org/10.1016/j.cedpsych.2003.09.002

Bakharia, A., Corrin, L., de Barba, P., Kennedy, G., Gašević, D., Mulder, R., Williams, D., Dawson, S., & Lockyer, L. (2016). A conceptual framework linking learning design with learning analytics. In D. Gašević, G. Lynch, S. Dawson, H. Drachsler, & C. Penstein Rosé (Eds.), *Proceedings of the sixth international conference on Learning Analytics & Knowledge - LAK '16* (pp. 329–338). ACM Press. https://doi.org/10.1145/2883851.2883944

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Brandl, L., Richters, C., Radkowitsch, A., Obersteiner, A., Fischer, M. R., Schmidmaier, R., Fischer, F., & Stadler, M. (2021). Simulation-based learning of complex skills: Predicting performance with theoretically derived process features. *Psychological Test and Assessment Modeling*, *63*(4), 542–560 https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2021-4/PTAM__4-2021_6_kor.pdf

Breakspear, S. (2012). *The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance* (OECD Education Working Papers No. 71). https://doi.org/10.1787/5k9fdfqffr28-en

Brooks, C., Baker, R., & Andres, J. M. L. (2015). Infrastructure for replication in learning analytics. In *Nature*. Advance Online Publicatio. https://doi.org/10.1038/nature.2015.17433

Care, E., Griffin, P., & McGaw, B. (2012). *Assessment and teaching of 21st century skills*. Springer.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: What will it take? *Theory Into Practice*, *42*(1), 75–83. https://doi.org/10.1353/tip.2003.0002

Dawson, S., Joksimovic, S., Poquet, O., & Siemens, G. (2019). Increasing the impact of learning analytics. In *Proceedings of the 9th international conference on Learning Analytics & Knowledge* (pp. 446–455). ACM. https://doi.org/10.1145/3303772.3303784

Fischer, F., Kollar, I., Ufer, S., Sodian, B., Hussmann, H., Pekrun, R., Neuhaus, B., Dorner, B., Pankofer, S., Fischer, M., Strijbos, J.-W., Heene, M., & Eberle, J. (2014). Scientific reasoning and argumentation: Advancing an interdisciplinary research agenda in education. *Frontline Learning Research*, *2*(3), 28–45. https://doi.org/10.14786/flr.v2i2.96

Gašević, D., Dawson, S., Rogers, T., & Gasevic, D. (2016). Learning analytics should not promote one size fits all: The effects of instructional conditions in predicting academic success. *The Internet and Higher Education*, *28*, 68–84. https://doi.org/10.1016/j.iheduc.2015.10.002

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analysing log file data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Methodology of educational measurement and assessment. Large-scale cognitive assessment* (pp. 239–269). Springer International Publishing. https://doi.org/10.1007/978-3-030-47515-4_10

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor: On validating the interpretation of process indicators based on log data. *Large-scale Assessments in Education*, *9*(1), 1–25. https://doi.org/10.1186/s40536-021-00113-5

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Methodology of educational measurement and assessment. Competence assessment in education* (pp. 407–425). Springer International Publishing. https://doi.org/10.1007/978-3-319-50030-0_24

Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263. https://doi.org/10.1016/j.compedu.2018.07.013

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, *61*, 36–46. https://doi.org/10.1016/j.chb.2016.02.095

Greiff, S., Wüstenberg, S., & Avvisati, F. (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*, *91*, 92–105. https://doi.org/10.1016/j.compedu.2015.10.018

Gür, B. S., Çelik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1–21. https://doi.org/10.1080/02680939.2011.595509

Hahnel, C., Ramalingam, D., Kroehne, U., & Goldhammer, F. (2023). Patterns of reading behaviour in digital hypertext environments. *Journal of Computer Assisted Learning*, *39*(3), 737–750. https://doi.org/10.1111/jcal.12709

Havnes, A., & Prøitz, T. S. (2016). Why use learning outcomes in higher education? Exploring the grounds for academic resistance and reclaiming the value of unexpected learning. *Educational Assessment, Evaluation and Accountability*, *28*(3), 205–223. https://doi.org/10.1007/s11092-016-9243-z

He, Q., Borgonovi, F., & Paccagnella, M. (2021). Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks. *Computers & Education*, *166*, 104170. https://doi.org/10.1016/j.compedu.2021.104170

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Advances in higher education and professional development (AHEPD) book series. Handbook of research on technology tools for real-world skill development* (pp. 750–777). Information Science Reference, an imprint of IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum or culture. *Educational Psychology*, *36*(2), 378–390. https://doi.org/10.1080/01443410.2014.946890

Ihantola, P., Vihavainen, A., Ahadi, A., Butler, M., Börstler, J., Edwards, S. H., Isohanni, E., Korhonen, A., Petersen, A., Rivers, K., Rubio, M. Á., Sheard, J., Skupas, B., Spacco, J., Szabo, C., & Toll, D. (2015). Educational data mining and learning analytics in programming. In N. Ragonis & P. Kinnunen (Eds.), *Proceedings of the 2015 ITiCSE on working group reports* (pp. 41–63). ACM. https://doi.org/10.1145/2858796.2858798

Jiang, Y., Gong, T., Saldivia, L. E., Cayton-Hodges, G., & Agard, C. (2021). Using process data to understand problem-solving strategies and processes for drag-and-drop items in a large-scale mathematics assessment. *Large-scale Assessments in Education*, *9*(1), 1–31. https://doi.org/10.1186/s40536-021-00095-4

Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *The NCME applications of educational measurement and assessment book series. Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). Routledge Taylor & Francis Group. https://doi.org/10.4324/9781315708591-2

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*(2), 527–563. https://doi.org/10.1007/s41237-018-0063-y

Kuhlmann, C., & Tillmann, K.-J. (2009). Mehr Ganztagsschulen als Konsequenz aus PISA? Bildungspolitische Diskurse und Entwicklungen in den Jahren 2000 bis 2003. In F.-U. Kolbe, S. Reh, T.-S. Idel, B. Fritzsche, & K. Rabenstein (Eds.), *Ganztagsschule als symbolische Konstruktion* (pp. 23–45). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91354-4_2

Li, H., Kim, M. K., & Xiong, Y. (2020). Individual learning vs. interactive learning: A cognitive diagnostic analysis of MOOC Students' learning behaviors. *American Journal of Distance Education*, *34*(2), 121–136. https://doi.org/10.1080/08923647.2019.1697027

Lockyer, L., Heathcote, E., & Dawson, S. (2013). Informing pedagogical action: Aligning learning analytics with learning design. *American Behavioral Scientist*, *57*(10), 1439–1459. https://doi.org/10.1177/0002764213479367

Lotz, C., Scherer, R., Greiff, S., & Sparfeldt, J. R. (2017). Intelligence in action – Effective strategic behaviors while solving complex problems. *Intelligence*, *64*, 98–112. https://doi.org/10.1016/j.intell.2017.08.002

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Mislevy, R. J. (2019). Advances in measurement and cognition. *The Annals of the American Academy of Political and Social Science*, *683*(1), 164–182. https://doi.org/10.1177/0002716219843816

OECD. (2010). *PISA computer-based assessment of student skills in science*. OECD Publishing. https://doi.org/10.1787/9789264082038-en

OECD. (2012). *PISA 2009 technical report*. OECD Publishing. https://doi.org/10.1787/9789264167872-en

OECD. (2013). *PISA 2012 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/9789264190511-en

OECD. (2016). *PISA in focus* (Vol. 67). Organisation for Economic Co-Operation and Development (OECD). https://doi.org/10.1787/aa9237e6-en

OECD. (2017). *PISA 2015 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*(1), 1–21. https://doi.org/10.1080/15305058.2013.825265

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science (New York, N.Y.)*, *372*(6540), 338–340. https://doi.org/10.1126/science.abd3300

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (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*, *12*, 579128. https://doi.org/10.3389/fpsyg.2021.579128

Rutkowski, L., & Rutkowski, D. (2016). A call for a more measured approach to reporting and interpreting PISA results. *Educational Researcher*, *45*(4), 252–257. https://doi.org/10.3102/0013189X16649961

Rutkowski, L., von Davier, M., & Rutkowski, D. (2013). *Handbook of international large-scale assessment*. Chapman and Hall/CRC. https://doi.org/10.1201/b16061

Skedsmo, G., & Huber, S. G. (2017). Policies and practices related to student assessment and learning outcomes—Combining different purposes and ideals. *Educational Assessment, Evaluation and Accountability*, *29*(3), 225–228. https://doi.org/10.1007/s11092-017-9268-y

Stadler, M., Becker, N., Gödker, M., Leutner, D., & Greiff, S. (2015). Complex problem solving and intelligence: A meta-analysis. *Intelligence*, *53*, 92–101. https://doi.org/10.1016/j.intell.2015.09.005

Stadler, M., Fischer, F., & Greiff, S. (2019). Taking a closer look: An exploratory analysis of successful and unsuccessful strategy use in complex problems. *Frontiers in Psychology*, *10*, 777. https://doi.org/10.3389/fpsyg.2019.00777

Stadler, M., Herborn, K., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: An investigation of the validity of the PISA 2015 CPS tasks. *Computers & Education*, *157*, 103964. https://doi.org/10.1016/j.compedu.2020.103964

Stadler, M., Hofer, S., & Greiff, S. (2020). First among equals: Log data indicates ability differences despite equal scores. *Computers in Human Behavior*, *111*, 106442. https://doi.org/10.1016/j.chb.2020.106442

Stadler, M., Radkowitsch, A., Schmidmaier, R., Fischer, M. R., & Fischer, F. (2020). Take your time: Invariance of time- on-task in problem solving tasks across expertise levels. *Psychological Test and Assessment Modeling*, *65*(4), 517–525.

Thille, C., Kizilee, R. F., Piech, C., Halawa, S. A., & Greene, D. K. (2014). The future of data–enriched assessment. *Research & Practice in Assessment*, *9*, 5–16.

Ulitzsch, E., He, Q., & Pohl, S. (2021). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, *47*(1), 3–35. https://doi.org/10.3102/10769986211010467

United Nations Educational, Scientific and Cultural Organization. (2015). *Education 2030 Incheon declaration and framework for action: Towards inclusive and equitable quality education and lifelong learning for all*. https://unesdoc.unesco.org/ark:/48223/pf0000245656

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(3), 324–343. https://doi.org/10.1080/0969594X.2014.999024

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, *44*(6), 671–705. https://doi.org/10.3102/1076998619881789

Wagemaker, H. (2014). *International large-scale assessments: From research to policy. Handbook of international large-scale assessment* (pp. 11–36). Background, Technical Issues, and Methods of Data Analysis.

Winthrop, R., & Simons, K. A. (2013). Can international large-scale assessments inform a global learning goal? Insights from the learning metrics task force. *Research in Comparative and International Education*, *8*(3), 279–295. https://doi.org/10.2304/rcie.2013.8.3.279

Wolff, A., Zdrahal, Z., Nikolov, A., & Pantucek, M. (2013). Improving retention. In D. Suthers, K. Verbert, E. Duval, & X. Ochoa (Eds.), *Proceedings of the third international conference on learning analytics and knowledge - LAK '13* (p. 145). ACM Press. https://doi.org/10.1145/2460296.2460324

Xiao, Y., He, Q., Veldkamp, B., & Liu, H. (2021). Exploring latent states of problem-solving competence using hidden Markov model on process data. *Journal of Computer Assisted Learning*, *37*(5), 1232–1247. https://doi.org/10.1111/jcal.12559

How to cite this article: Stadler, M., Brandl, L., & Greiff, S. (2023). 20 years of interactive tasks in large-scale assessments: Process data as a way towards sustainable change? *Journal of Computer Assisted Learning*, *39*(6), 1852–1859. https://doi.org/10.1111/jcal.12847