Organisation for Economic Co-operation and Development

EDU/WKP(2023)1

Unclassified **English text only**

20 January 2023

DIRECTORATE FOR EDUCATION AND SKILLS

The Uses of Process Data in Large-Scale Educational Assessments

OECD Education Working Paper No. 286

Professor Bryan Maddox, Digital Education Futures Initiative (DEFI), Hughes Hall, University of Cambridge & Assessment MicroAnalytics Ltd.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Bryan Maddox: bryan.maddox@hughes.cam.ac.uk

JT03511084

OECD EDUCATION WORKING PAPERS SERIES

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Acknowledgements

This working paper has been produced with the support of the Directorate for Education and Skills at the OECD. In particular, I would like to acknowledge the feedback and support of Mario Piacentini, Ava Guez, François Keslair and Yuri Belfali in its drafting and preparation, as well as Rachel Linden for the editorial work.

Abstract

The digital transition in educational testing has introduced many new opportunities for technology to enhance large-scale assessments. These include the potential to collect and use log data on test-taker response processes routinely, and on a large scale. Process data has long been recognised as a valuable source of validation evidence in assessments. However, it is now being used for multiple purposes across the assessment cycle. Process data is being deliberately captured and used in large-scale, standardized assessments – moving from viewing it as a 'by-product' of digital assessment, to its use 'by design' to extend understanding of test-taker performance and engagement. While these techniques offer significant benefits, they also require appropriate validation practices to ensure that their use supports reliable inferences and does not introduce unintended negative consequences.

Table of contents

Figures

1. Introduction

The digital transition is having profound implications for large-scale educational assessments, not only on the mode of delivery, with increased personalisation, accessibility, interaction, and user engagement, but also because of the potential of digital assessments to improve the way that data is collected and used (Goldhammer, Scherer and Greiff, 2020[1]; Jiao, He and Veldkamp, 2021[2]). In paper-based modes of assessments, response processes that take place between stimulus and response largely go unobserved. On-screen, digital assessments let us dig much deeper into student performance because it enables us to routinely capture and analyse the clickstream (log data) on student interactions with the keyboard and mouse. The use of process data also supports in-depth probes into student performance on test items, for example with the use of eye tracking, video and screen capture, and with physiological measures. Whereas think-aloud protocols are usually available for a small number of participants, digital log data is routinely collected for the entire tested population. As a result, data on assessment response processes has started to be exploited in many ways across the assessment cycle, from the design and field testing of test items to quality assurance, enhancing the ways that we understand test engagement and performance, and how we validate the interpretation and use of assessment results.

The collection and analysis of data on student response processes has advanced quickly in recent years and involves important new areas of activity. The first is the post-hoc analysis of the process data that is generated as a 'by-product' in large-scale assessment, for example, using log data on item response times and keystrokes (Goldhammer, Scherer and Greiff, 2020[1]; Ercikan, Guo and He, 2020[3]). Through the secondary analysis of such large-scale data, this first area focuses on the comparison of data on performance, motivation, and engagement within and across population groups and contexts. For example, rapid response times, alongside declining test performance may be associated with disengagement and guessing (Wise, 2020[4]). A second area of activity involves the analysis of process data from cognitive labs, field trial, and in-situ observations in order to provide timely interventions by improving test design, user experience, and construct validation (Kane and Mislevy, 2017[5]; Ercikan and Pellegrino, 2017[6]). For example, data from eye tracking studies and think aloud can provide evidence on how people understand and engage with item content (Lindner et al., 2018[7]). A final area of activity focuses on how the digital transition is impacting on test design. Digital assessments can capture and use data on the ways that people interact and engage with test items. 'Digital first' assessment designs anticipate and deliberately build-in the use of process data by-design, moving beyond the notion of process data as a by-product of assessment, to place it at the centre of test (Goldhammer et al., 2021[8]; Salles, Dos Santos and Keskpaik, 2020[9]; Burstein et al., 2021[10]). This can include the collection of clickstream data on response times and keystrokes, which can be used to gain deeper, more granular insights into test-taker performance and engagement, and to capture data on 'process-oriented' assessment constructs such as problem solving, interaction and collaboration.

The development of process-oriented digital assessment designs presents at the same time an opportunity and a challenge. The opportunity involves the scope to take advantage of the potentials of digital assessment to create highly engaging and interactive assessment tasks that reflect and take advantage of digital first designs. The challenge involves the task of establishing the validity, reliability, ethics and fairness of those uses of process data, to the extent that they can be routinely incorporated into the formal processes of large-scale assessments (Goldhammer and Zehner, 2017[11]; Kroehne and Goldhammer, 2018[12]; Goldhammer et al., 2021[8]; Han, Krieger and Greiff, 2021[13]; Murchan and Siddiq, 2021[14]). Those considerations are particularly important in international large-scale assessments, which must support the comparability of process data across diverse groups

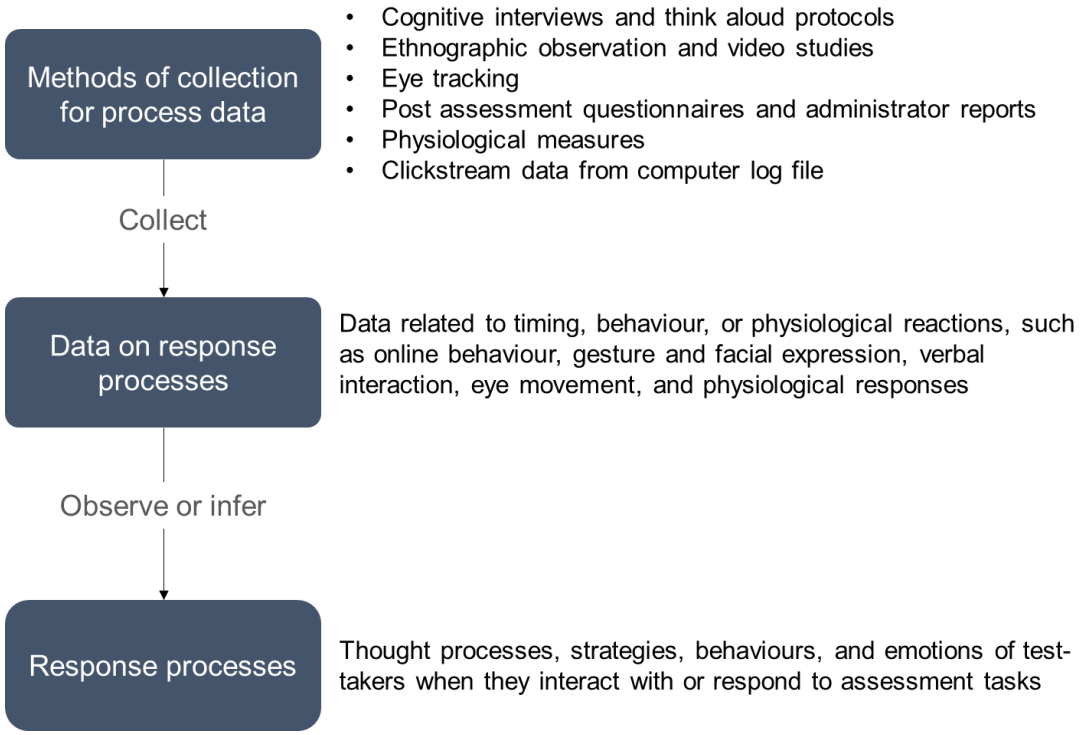
and contexts (Oliveri, Lawless and Mislevy, 2019[15]; Addey, Maddox and Zumbo, 2020[16]; Ercikan, Guo and He, 2020[3]).

This Working Paper is therefore intended to support the systematic use of process data in large-scale educational assessments. The paper begins with a discussion of the definitions of process data. It then describes the various uses of process data across the assessment cycle, and the challenge of validating the use of process data. The paper discusses how process data is used to generate improved understanding of test-taker performance and engagement. It concludes by highlighting the importance of developing appropriate ethical codes and frameworks.

2. Definitions

Process data concerns sources of information about assessment response processes and the testing situation that may be used to generate inferences about some characteristics of test performance (see Figure 1).

Figure 1: Collecting data on response processes



Response processes. Discussions of response processes typically highlight its multiple features and dimensions. For example, Ercikan and Pellegrino state that 'Response processes refer to the thought processes, strategies, approaches, and behaviours of examinees when they read, interpret and formulate solutions to assessment tasks.' (Ercikan and Pellegrino, 2017, p. 2[6]). Their description identifies diverse phenomena, including thought processes and strategies, as well as behaviours. Some of those may be readily inferred from response process data, while others suggest underlying processes. Hubley and Zumbo (2017[17]) take a similar approach when they define response processes as: 'the mechanisms that underlie what people do, think, or feel when interacting with, and

responding to, the item or task and are responsible for generating observed test score variation'. This definition expands response processes beyond the cognitive realm to include emotions, motivations and behaviours (Hubley and Zumbo, 2011, p. 2[18]). These definitions have a shared concern with the testing situation as a distinct domain of enquiry (McNamara and Roever, 2006[19]; Maddox, 2015[20]; Maddox and Zumbo, 2017[21]). That is, they aim to capture the distinctive ecology of the testing situation that may be lost in conventional 'product' data on test scores. In the context of digital, screen-based assessments, those features of the testing situation include human-computer interaction, User Experience and engagement with digital interfaces and digital tools. Data about these features can often be useful in helping to explain observed variation in test scores that may not be construct relevant, for example to provide evidence on how test scores interact with some wider characteristics of group membership, disabilities, socio-cultural and linguistic contexts (Zumbo, 2015[22]; Ercikan, Guo and He, 2020[3]). In that way, process data is an important source of information that can support work on test fairness and inclusion.

Data on response processes. There are multiple types of process data in assessment (e.g., timing, behaviour, physiological reactions), and each contributes to understanding of some aspect of how test takers engage with assessment tasks (Ercikan and Pellegrino, 2017[6]; Oranje et al., 2017[23]; Hubley and Zumbo, 2017[17]; Ercikan, Guo and He, 2020[3]). These include sources of data on online behaviour, gesture and facial expression, verbal interaction, eye movement, and physiological responses.

Methods of collection for process data. The techniques and methods used to collect data on response processes include:

- Cognitive interviews and think-aloud protocols (Pepper et al., 2018[24]; Padilla and Benitez, 2017[25]);
- Ethnographic observation and video studies (Maddox, 2014[26]; Maddox, 2017[27]; Maddox, 2018[28]; Maddox and Zumbo, 2017[21]; Maddox, Keslair and Jayrh, 2019[29]);
- Eye tracking (Oranje et al., 2017[23]; Lindner et al., 2017[30]; Lindner et al., 2018[7]; Maddox et al., 2018[31]);
- Post-assessment questionnaires and administrator reports (Eklöf and Knekta, 2017[32]; Eklöf and Hopfenbeck, 2019[33]; Hopfenbeck and Kjærnsli, 2016[34]);
- Physiological measures (Aryadoust, Foo and Ng, 2022[35]);
- and various uses of clickstream data from computer log files (i.e., keystrokes, mouse movements) to investigate and draw inferences about response times and interactions (Wise and Kong, 2005[36]; Wise, 2017[37]; Reis Costa et al., 2021[38]; Salles, Dos Santos and Keskpaik, 2020[9]; Michaelides, Ivanova and Nicolaou, 2020[39]; Goldhammer et al., 2021[40]; Deribo, Goldhammer and Kroehne, 2022[41]).

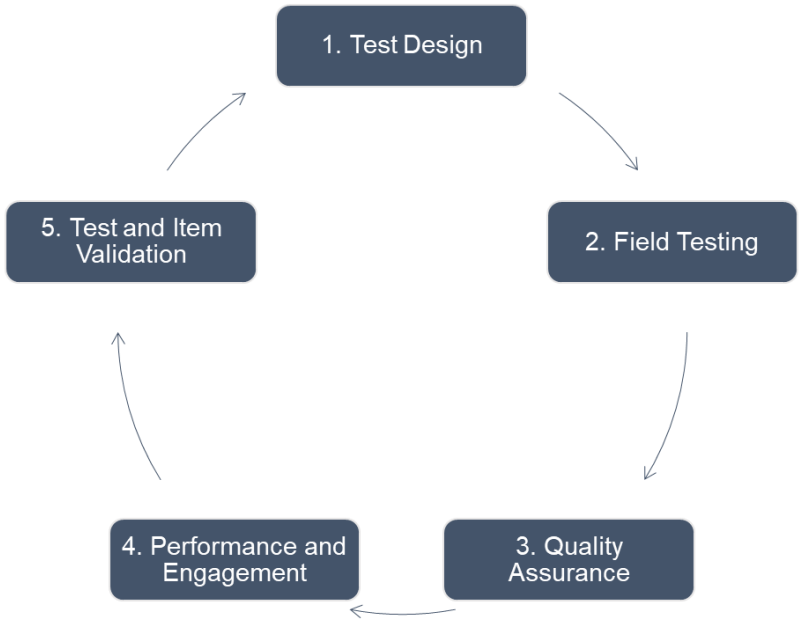
The specifications of process data, methodological approaches, and validation procedures therefore vary according to the methods of collection and the types of use. The field is characterised as one of methodological pluralism (Ercikan and Pellegrino, 2017[6]; Hubley and Zumbo, 2017[17]). However, the digital transition supports the automated collection of various sources of process data, with sufficient temporal and spatial resolution and granularity to enable the analysis of response processes within items (Goldhammer, Scherer and Greiff, 2020[1]; Lindner and Greiff, 2021[42]). Whether it involves process data 'by design', or retrospective data mining, the automated collection and use of big data on assessment response processes implies greater requirements for formalisation and transparency of methods, validity procedures, clarity of purpose, and institutional development of data infrastructures and design architectures (Kroehne and Goldhammer,

2018[12]; Gulson and Sellar, 2019[43]; Goldhammer et al., 2021[8]; Kespaik, Dos Santos and Salles, 2021[44]; Piattoeva and Vasileva, 2021[45]).

3. The uses of process data

It has long been recognised that information about the way that respondents engage with test items, and the strategies they use to formulate their answers, can provide important feedback to test designers, and for the validation of assessment constructs (Cronbach and Meehl, 1955[46]; Lennon, 1956[47]; Messick, 1989[48]; Embretson, 1984[49]). At its most simple level, requests for test takers to 'show their working' can provide useful insights into how they reach their answers. Similarly, in test design and validation, 'probes' into response processes such as think-aloud protocols have long been recognised as a source of valuable insights about how test takers understand tasks, and how they reach their submitted answers (Messick, 1989[48]). Historical interest in process data has focused on its use in test validation. That is reflected in the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014[50]). However, the contemporary uses of process data also extend beyond test validation. The reason is the coming together of the affordances of digital assessments, recent advances in data science and computational psychometrics, that have enabled large-scale analysis of the detail of item responding, most notably via the application of log files in computer-based assessments (Goldhammer, Scherer and Greiff, 2020[1]). Digital process data is collected on all students as a feature of digital assessments. Furthermore, the scope for in-depth analysis of the cognitive and affective dimensions of assessment response processes have significantly enhanced with the application of digital data from sources such as eye tracking, web-cam video and audio feeds. As a result, there is an explosion in the uses of process data, for multiple purposes, across the assessment cycle (see Figure 2).

Figure 2: Uses of process data across the assessment cycle



While process data, particularly log data, are viewed as a valuable resource for post-hoc analysis, the uses of process data also point to an extensive integration of process data across the entire assessment cycle. This highlights a goal oriented, pragmatic approach to collecting and using process data, to improve many aspects of assessment design, test quality, and validation.

Figure 2 indicates the different uses of process data across the assessment cycle. While each use has relevance to validation, each use has its own distinctive rationale and purpose, which informs the specifications, scale, sources, and use of data involved (as discussed below). For example, the design and field testing of tests and items might involve in-depth observational studies with relatively small sample sizes that can rapidly inform improvements to test design and administration. In contrast, the use of large-scale uses of log data to enhance the analysis of test-taker performance and engagement usually require larger-scale samples. Each of these uses of process data requires its own validation.

1. **Test Design:** The uses of process data have become integral to iterative processes of test design, to ensure that the response processes and user experience, test accessibility, support the rationales for test design and use (Kane and Mislevy, 2017[5]). Process data from sources such as eye tracking and think-aloud and cognitive interviews, and studies of user experience, provide in-depth data that are used to identify anomalies in response processes linked to unanticipated threats to test performance and validity, and to make iterative improvements to test design (Gorin, 2006[51]; Bax and Chan, 2019[52]; Oranje et al., 2017[23]; Maddox et al., 2018[31]; Padilla and Leighton, 2017[53]; Yaneva et al., 2021[54]).
2. **Field Testing:** Process data from field tests provide important sources of information to inform test design, test administration and validation by providing evidence about the ways that diverse groups of test takers, socio-economic, cultural and linguistic contexts understand and receive test item content, and navigate within and between test items. This is especially important in International, Large-Scale Assessments, where differences in the testing situation, test administration, and the wider test ecology may introduce sources of bias and variation in test data (Zumbo et al., 2015[55]; Li, Hunter and Bialo, 2021[56]). Sources of information in field testing include ethnographic observations, post-assessment interviews, and eye tracking, as well as log data on keystrokes and response times (Oliveri, Lawless and Mislevy, 2019[15]; Addey, Maddox and Zumbo, 2020[16]; Maddox and Zumbo, 2017[21]).
3. **Quality Assurance:** Process data plays a central role in quality assurance, and to identify unexpected anomalies in test administration and reception that may indicate threats to data quality and validity including variation in test reception within or across assessment systems and contexts. The sources of process data in quality assurance are varied, for example, administrator and scorer reports, ethnographic observations, GPS data on the movement of test administrators, and data forensics from log file information on keystrokes and response times (Yamamoto and Lennon, 2018[57]; Wise, Kuhfeld and Soland, 2019[58]; Maddox, 2014[26]; Maddox, 2017[27]; Maddox, 2015[20]; Maddox, Keslair and Jayrh, 2019[29]; Maddox et al., 2015[59]). Quality assurance data such as video and audio recordings also provide a rich source of evidence on variation in test administration, and the presence of cheating and score fabrication, as well as wider threats to test quality and reliability associated with differences in test reception across cultures and contexts (Yamamoto and Lennon, 2018[57]).
4. **Engagement and Performance:** Process data has become extensively used to generate insights and evidence on aspects of student engagement and performance (Goldhammer, Scherer and Greiff, 2020[1]; Jiao, He and Veldkamp, 2021[2]; Lundgren and Eklöf, 2020[60]). This accounts for the rapid growth in interest in the uses of process data in large-scale assessments, with significant implications for test rationales, design and validation, and for

emergent infrastructures and techniques for data analysis, models and interpretation. These techniques initially treated process data from computer log files as a 'by-product' of assessment, but that is rapidly being replaced by more deliberate integration of such data into the item design and rationales for the measurement of student engagement and performance. This includes the use of process data 'by design' that anticipates and integrates the use of keystroke data into test items, to complement (or even replace) the conventional test score or product. This can be done by integrating process data in real time to inform the calculation of test scores, or to inform routing decisions on computer-adaptive (CAT) designs, or retrospectively, using computational techniques and data mining to enhance understanding of some aspect of test performance (e.g., student ability, engagement, equity, fairness).

5. **Test and Item Validation:** Process data provide a rich source of information about what happens between 'stimulus and response', of the type that is unobserved (or black boxed) in conventional data on test scores. For example, it can help to explain sources of variation in test-taker performance, and to support the valid interpretation of test scores. In this way, process data can provide important complementary sources of evidence alongside conventional psychometric data on test scores and item characteristics (Ercikan, Guo and He, 2020[3]). The inclusion of process data can therefore be considered within a holistic framework of test validation (Zumbo, Maddox and Care, 2023[61]).

Within the applications of process data described above, the themes 5 above are of major concern to assessment organisations working in large-scale assessments, namely, the use of process data to enhance measures of test engagement, and its use to enhance understanding of student performance. These themes merit some further discussion.

4. Measures of test engagement

Test-taker disengagement is a particular concern in large-scale, low-stakes assessments. The general concern is that disengaged test taking is a threat to measurement validity including the possibility of declining performance over the duration of the test. That has stimulated considerable attention in the research literature on 'disengaged rapid guessing'. Differences in test-taker engagement in large-scale assessments have been shown to be associated with large variation in test scores with associated variation by country and by gender (Gneezy et al., 2019[62]; Ranger, Kuhn and Pohl, 2021[63]). The implication being not only that it is necessary to systematically capture variation in engagement in order to make valid interpretations of test scores, but also, that improvements in test design, test enjoyment, accessibility and user experience, are necessary to reduce the scale of disengagement (Burstein et al., 2021[10]; Care and Maddox, 2021[64]).

There are multiple measures of test engagement. These include questionnaires and surveys such as the Programme for International Student Assessment (PISA) 'effort thermometer' and 'perseverance index' (Eklöf and Knekta, 2017[32]; Eklöf and Hopfenbeck, 2019[33]), the use of eye tracking studies (Oranje et al., 2017[23]; Maddox, 2018[28]), and rapidly expanding use of log data on item response times and keystrokes (Goldhammer, Martens and Lüdtke, 2017[65]; Wise, 2017[37]; Wise, 2019[66]; Wise, Kuhfeld and Soland, 2019[58]; Lee and Jia, 2014[67]; Gneezy et al., 2019[62]; Kroehne, Deribo and Goldhammer, 2020[68]; Wise, 2020[69]).

Log data on 'rapid guessing behaviours' have generated attention in studies of engagement because they suggest little overlap of disengaged test taking behaviours and response times associated with legitimate 'solution behaviours' (Wise, 2019[66]). The identification of rapid guessing is therefore a valuable source of retrospective data on respondent

disengagement in large-scale assessments, particularly as it can be used to identify problems of item fit that may apply to particular groups or contexts, of sources of variation in engagement associated with local test administration.

The research literature has established multiple methods to identify disengaged rapid guessing based on item response times in large-scale assessments (Wise, 2019[66]). Those methods provide various attempts to estimate response time 'thresholds' associated with rapid guessing. There are, however, potential threats to the validity of data on rapid guessing as a measure of test engagement. Notably, rapid guessing behaviours are found to vary across different groups, they relate to the mode of assessment, and the design features of test items (Wise, 2019[66]; Kroehne, Deribo and Goldhammer, 2020[68]). That suggests the need for caution about the interpretation of rapid response behaviours as test-taker disengagement, as its presence may indicate wider sources of variation.

A second, and perhaps more profound threat to the use of rapid response behaviours as a measure of disengagement, is that many disengaged behaviours are associated with response times within the normal distribution, where there is a mixture of guessing behaviours and solution behaviours (Wise, 2019[66]). Furthermore, test takers may exhibit a mixture of engaged and disengaged behaviours – such as making engaged attempts to answer an item, before becoming disengaged and guessing or not submitting an answer (Maddox, 2017[27]). The purposeful capture of granular data on test taking behaviours within items, for example in process-oriented items (where respondents are expected to engage with different tools and resources) therefore promises improved accuracy in the detection and modelling of different types test item disengagement. This, for example may enable a greater differentiation between generalised disengagement and fatigue, and disengagement that relates to the particular content and features of test items.

5. Measures of performance

Measures of respondent performance are increasingly used to supplement, enhance or replace conventional test scores (Mislevy et al., 2014[70]; von Davier et al., 2017[71]; Rojas et al., 2021[72]; Stoeffler et al., 2020[73]). This includes log data on item response times (Li, Banerjee and Zumbo, 2017[74]; Reis Costa et al., 2021[38]; Ercikan, Guo and He, 2020[3]; Deribo, Goldhammer and Kroehne, 2022[75]) and clickstream data about response processes within the item – including type and number of 'actions' and 'events' that can be used to infer the strategies that respondents use in tacking items (Salles, Dos Santos and Keskpaik, 2020[9]; Goldhammer et al., 2021[8]; Ercikan, Guo and He, 2020[3]). While item response times provide a signal about variation in response processes and their relationship with test scores, response data collected on performance within items, i.e., from clickstream, eye tracking and think aloud provides a richer source of information to inform and extend measures of item performance. This can be used alongside data on response times.

A distinction can be made between the use of response process data to enhance and extend understanding of performance in more established, conventional test items such as the domains of mathematics, language and science, where there is a clear answer or 'product', and those of innovative domains such as creativity and problem solving, where the construct being assessed is inherently process oriented, and where the item is designed to measure the way that respondents engage with certain tasks and challenges within the item, such as accessing and using information, and interaction with non-human agents (avatars, chat bots) or human participants.

An example of the uses of process data to enhance analysis of respondent performance in conventional assessment domains is the use of process data supported by large-scale data

mining in French secondary school mathematics assessments with the aim of enhancing the interpretation of assessment data, and to inform improvements to teaching practice (Salles, Dos Santos and Keskpaik, 2020[9]; Kespaik, Dos Santos and Salles, 2021[44]). The team at the Department of Evaluation (DEPP) at the French Ministry of Education analysed log data with data mining techniques to model the ways that school students tackle interactive digital mathematics tasks. Their analysis was supported by a didactic analysis, and later with data from eye tracking and retrospective think aloud to refine their interpretive models.

A contrasting case is the design of digital first assessments in innovative domains. Those assessments, for example, of problem solving, creativity and computational skills are designed from the outset with the assumption that process data will provide the primary source of information on respondent performance (Fiore, Graesser and Greiff, 2018[76]; He, Borgonovi and Paccagnella, 2021[77]; Han, Krieger and Greiff, 2021[13]; Stoeffler et al., 2020[73]; Ercikan, Guo and He, 2020[3]). In those cases, 'innovation' relates not only to the assessment domain, but to radically 'disruptive' assumptions about the way that assessment is conducted and validated (Stoeffler et al., 2020[73]; Wyatt-Smith, Lingard and Heck, 2021[78]). In the process-oriented assessment of innovative domains, measures of performance are designed into the items through the collection of clickstream events and timestamps, and data on verbal interaction data on how people engage with online tools and resources (Andrade et al., 2019[79]).

In the deliberate uses of process data by design to measure student performance, upstream design processes require considerable investment in iterative, evidenced design and validation to ensure user experience and responses support process model assumptions (Kane and Mislevy, 2017[5]). That design process is quite different to retrospective data mining of process, for example, that use time stamp data or information on within item log events to yield large-scale supplementary data on student performance from more conventional, product-oriented assessment designs (e.g., (Ercikan, Guo and He, 2020[3]; Reis Costa et al., 2021[38])). However, whether it is for process data by design, or in retrospective data mining, any use of process data to make performance related inferences require validation arguments, models and evidence (Goldhammer et al., 2021[8]).

The greater the reliance on process data for measures of performance, the higher the demands are for appropriate validation, as it is not always clear that variation in process-oriented measures are construct relevant within and across user groups and cultural contexts (Zumbo, Maddox and Care, 2023[61]; Ercikan, Guo and He, 2020[3]). Furthermore, the design of process related items should be able to demonstrate the robustness of process models at the level of individual respondents (as it is in product-oriented designs), rather than at the level of large-scale aggregations. That requires in-depth work in item design and pilot processes (such as combined work on think aloud and eye tracking and log data analytics).

6. Validating the uses of process data

Like the use of conventional test scores, any interpretation and use of process data requires validation, informed by appropriate arguments, warrants and evidence, and considerations of the reliability, fairness and consequences of using process data (Kane and Mislevy, 2017[5]; Goldhammer et al., 2021[8]). Any use of process data therefore needs to be accompanied by those arguments and evidence, particularly to ensure accountability to the participants of assessment and those who will make use of and be impacted by the intended uses of process data.

Although process data from digital and automated methods are sometimes thought to be 'collected' – for example, as a by-product of computer-based testing, they are nevertheless constructed phenomena, that cannot escape from the methodological assumptions, treatments and the application of theoretical models (Gulson and Sellar, 2019[43]; Goldhammer et al., 2021[8]; Mislevy, 2018[80]). The ubiquitous presence of log data, the granularity and resolution of eye tracking, or the technical presence of physiological measures does not negate the requirement for interpretation and use arguments, and appropriate validation. There are several sources of threats to the valid interpretation and use of process data in large-scale assessments as follows.

6.1. Partiality

Since assessment response processes are multi-dimensional, any single method used for the collection of process data is vulnerable to threats to validity associated with methodological bias and partiality (Lee and Haberman, 2015[81]). Those risks, and the need for appropriate triangulation using process data from other sources are particularly important for uses of process data to generate inferences about respondent performance and engagement (Li, Banerjee and Zumbo, 2017[74]; Goldhammer et al., 2021[8]). Triangulation with wider sources of process data such as eye tracking and think aloud, and video studies is therefore valuable to validate the interpretation use of process data. Log data is especially vulnerable to threats to validity because of its partiality, as it only captures log events from respondent interactions with the keyboard and mouse. It does not, for example, capture data on respondent behaviour during 'idle time' (away from keyboard), between the clicks, such as patterns of reading behaviour, or off-screen activity such as use of pen and paper (Maddox, 2017[27]; Salles, Dos Santos and Keskpaik, 2020[9]).

6.2. Theoretical Constructs

Sensor based data on response processes are not necessarily related to the assessment constructs in ways that are immediately obvious. In most cases those interpretations are mediated by theoretical constructs that help to account for underlying phenomena such as latent attributes of personality, cognition, motivation etc. The choice and use of theoretical constructs, and its perceived association with process data therefore requires justification and validation (Goldhammer et al., 2021[8]; Ercikan, Guo and He, 2020[3]; Jiao, He and Veldkamp, 2021[2]; Maddox, 2017[27]; Hahnel et al., 2019[82]). The decision to draw from underling theory is not in-itself adequate evidence of its validity.

6.3. Diversity, Equity, and Inclusion

The valid interpretation and use of data on response processes need to consider the sources of diversity in the tested population that could lead to unintended, construct irrelevant sources of variation in process data. This may include characteristics that impact on the response processes of test takers, including neurodiversity, disability, and linguistic, and cultural diversity. For example, students who have disabilities such as Autistic Spectrum Disorder (ASD), or physical disabilities that impact on motor movement might influence their response times and keystrokes in ways that are not relevant to the assessment constructs (Zumbo, Maddox and Care, 2023[61]). Similarly, researchers have observed 'Differential Response Times' (DRT) associated with linguistic diversity that are similar to Differential Item Functioning (DIF) (Ercikan, Guo and He, 2020[3]). This suggests that process data – for example on response times, might be used to investigate and support agendas on diversity, equity and inclusion.

6.4. Ethics & Consequences

As we have seen, digital assessments, and the uses of process data profoundly shape the types of data that is collected, and how it is stored and used. Sources of process data are considerably more invasive than in conventional, 'product' oriented assessments since they capture and use intimate behaviours as respondents formulate solutions to assessment tasks. They may also capture audio and video recordings of test takers. This has significant implications for data ethics (including informed consent and data security), the need to demonstrate fairness and transparency, and for establishing and maintaining public trust (Murchan and Siddiq, 2021[14]; Southgate, 2021[83]). As Murchan and Siddiq (2021[14]) have argued, the uses of process data in large-scale educational assessments require urgent work to establish appropriate ethical frameworks and protocols to regulate its use.

Teacher and test-taker perceptions about, or concerns with the way that process data is collected and used in assessments may undermine the perceived validity of assessments (i.e., face validity), or create 'washback', and unintended negative consequences in terms of how they prepare to take assessments (Sellar et al., 2019[84]; Gulson and Sellar, 2019[43]; Johnson and Shaw, 2019[85]; Knox, Williamson and Bayne, 2020[86]).

7. Conclusion

The digital transition has introduced many opportunities for technology to enhance and transform the work of large-scale assessments. Of those, the routine capture and use of data on response processes has the potential to significantly improve the quality and reliability of large-scale assessments. As we have seen, that includes the uses of process data across the assessment cycle, with particular potential to improve the quality and volume of information on student performance and engagement.

The use of process data 'by design' involves a step change for large-scale assessment. Many of the initial uses of process data were opportunistic retrospective, treating process data as a convenient by-product of computer-based assessment. Those approaches have generated new methods and insights that are now evident in an extensive body of research publications. However, ad-hoc and retrospective approaches tend to be time consuming – their results are rarely prepared in time to be published alongside conventional test score data or in technical reports. In contrast, the deliberate and purposive collection of process data 'by design' creates opportunities for its systematic use as a key element in assessment programmes to enhance test quality and validity. They also enable computational methods to be used to improve the assessment of distinctively 'process' related constructs such as problem solving and interaction with the digital environment.

The uses of process data also present some challenges and risks. Those include the need to establish suitable digital infrastructures, to develop appropriate design arguments and validation practices that can be integrated into mainstream assessment practice. A key indicator for success would be the extent to which those arguments, technical procedures and data are represented in framework documents, reports of assessment results, and in technical reports.

To conclude, we can identify the following high-level recommendations about the uses of process data in large-scale assessments. Firstly, the uses of process data 'by design' (rather than as an ad-hoc basis) should be integrated into test constructs, item design, data infrastructures, and validation processes. Second, the potential for unintended negative consequences should be considered and researched, as well as the opportunities for the uses of process data to support agendas of diversity, equity and inclusion. Finally, reports on the

collection, interpretation and use of process data should be fully integrated into the reporting and publications of testing organisations. In that context, public facing communication on the rationales and arguments for the uses of process data are required to build stakeholder trust and understanding.

References

Addey, C., B. Maddox and B. Zumbo (2020), "Assembled validity: rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs)", *Assessment in Education: Principles, Policy and Practice*, Vol. 27/6, https://doi.org/10.1080/0969594X.2020.1843136. [16]

AERA, APA and NCME (2014), *Standards for Educational and Psychological Testing, 2014 Edition*. [50]

Andrade, A. et al. (2019), "Multimodal Interaction Analysis for the Assessment of Problem-Solving Skills in a Collaborative Online Game", in Eagan, B., M. Misfeldt and A. Siebert-Evenstone (eds.), *Advances in Quantitative Ethnography First International Conference*, Springer, Madison, USA. [79]

Aryadoust, V., S. Foo and L. Ng (2022), "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, https://doi.org/10.1177/02655322211026876. [35]

Bax, S. and S. Chan (2019), "Using eye-tracking research to investigate language test validity and design", *System*, Vol. 83, https://doi.org/10.1016/j.system.2019.01.007. [52]

Burstein, J. et al. (2021), *A Theoretical Assessment Ecosystem for a Digital-First Assessment—The Duolingo English Test*, Duolingo Research Report DRR-21-04. [10]

Care, N. and B. Maddox (2021), *Improving Test Validity and Accessibility with Digital First Assessments*, White Paper, Duolingo. [64]

Cronbach, L. and P. Meehl (1955), "Construct validity in psychological tests", *Psychological Bulletin*, Vol. 52/4, https://doi.org/10.1037/h0040957. [46]

Deribo, T., F. Goldhammer and U. Kroehne (2022), "Changes in the Speed–Ability Relation Through Different Treatments of Rapid Guessing", *Educational and Psychological Measurement*, p. 001316442211094, https://doi.org/10.1177/00131644221109490. [41]

Deribo, T., F. Goldhammer and U. Kroehne (2022), "Changes in the Speed–Ability Relation Through Different Treatments of Rapid Guessing", *Educational and Psychological Measurement*, p. 001316442211094, https://doi.org/10.1177/00131644221109490. [75]

Eklöf, H. and T. Hopfenbeck (2019), "Self reported effort and motivation in the PISA test", in Maddox, B. (ed.), *International Large-Scale Assessments in Education*, Bloomsbury, London. [33]

Eklöf, H. and E. Knekta (2017), "Using large-scale educational data to test motivation theories: a synthesis of findings from Swedish studies on test-taking motivation", *International Journal of Quantitative Research in Education*, Vol. 4/1/2, https://doi.org/10.1504/ijqre.2017.086499. [32]

Embretson, S. (1984), "A general latent trait model for response processes", *Psychometrika*, Vol. 49, pp. 175-186. [49]

Ercikan, K., H. Guo and Q. He (2020), "Use of Response Process Data to Inform Group Comparisons and Fairness Research", *Educational Assessment*, Vol. 25/3, https://doi.org/10.1080/10627197.2020.1804353. [3]

Ercikan, K. and J. Pellegrino (2017), *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*, Routledge. [6]

Fiore, S., A. Graesser and S. Greiff (2018), *Collaborative problem-solving education for the twenty-first-century workforce*, https://doi.org/10.1038/s41562-018-0363-y. [76]

Gneezy, U. et al. (2019), "Measuring Success in Education: The Role of Effort on the Test Itself", *American Economic Review: Insights*, Vol. 1/3, https://doi.org/10.1257/aeri.20180633. [62]

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, p. 20, https://doi.org/10.1186/s40536-021-00113-5. [8]

Goldhammer, F. et al. (2021), "Controlling speed in component skills of reading improves the explanation of reading comprehension.", *Journal of Educational Psychology*, Vol. 113/5, pp. 861-878, https://doi.org/10.1037/edu0000655. [40]

Goldhammer, F., T. Martens and O. Lüdtke (2017), "Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics", *Large-Scale Assessments in Education*, Vol. 5/1, https://doi.org/10.1186/s40536-017-0051-9. [65]

Goldhammer, F., R. Scherer and S. Greiff (2020), *Editorial: Advancements in Technology-Based Assessment: Emerging Item Formats, Test Designs, and Data Sources*, https://doi.org/10.3389/fpsyg.2019.03047. [1]

Goldhammer, F. and F. Zehner (2017), *What to Make Of and How to Interpret Process Data*, https://doi.org/10.1080/15366367.2017.1411651. [11]

Gorin, J. (2006), "Test design with cognition in mind", *Educational Measurement: Issues and Practice*, Vol. 25/4, https://doi.org/10.1111/j.1745-3992.2006.00076.x. [51]

Gulson, K. and S. Sellar (2019), "Emerging data infrastructures and the new topologies of education policy", *Environment and Planning D: Society and Space*, Vol. 37/2, https://doi.org/10.1177/0263775818813144. [43]

Hahnel, C. et al. (2019), "Validating process variables of sourcing in an assessment of multiple document comprehension", *British Journal of Educational Psychology*, Vol. 89/3, https://doi.org/10.1111/bjep.12278. [82]

Han, A., F. Krieger and S. Greiff (2021), "Collaboration analytics need more comprehensive models and methods. An opinion paper", *Journal of Learning Analytics*, Vol. 8/1, https://doi.org/10.18608/JLA.2021.7288. [13]

He, Q., F. Borgonovi and M. Paccagnella (2021), "Leveraging process data to assess adults' problem-solving skills: Using sequence mining to identify behavioral patterns across digital tasks", *Computers and Education*, Vol. 166, https://doi.org/10.1016/j.compedu.2021.104170. [77]

Hopfenbeck, T. and M. Kjærnsli (2016), "Students' test motivation in PISA: the case of Norway", *The Curriculum Journal*, Vol. 27/3, pp. 406-422, https://doi.org/10.1080/09585176.2016.1156004. [34]

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. [17]

Hubley, A. and B. Zumbo (2011), *Validity and the Consequences of Test Interpretation and Use*, https://doi.org/10.1007/s11205-011-9843-4. [18]

Jiao, H., Q. He and B. Veldkamp (2021), "Editorial: Process Data in Educational and Psychological Measurement", *Frontiers in Psychology*, Vol. 12, https://doi.org/10.3389/fpsyg.2021.793399. [2]

Johnson, M. and S. Shaw (2019), "What is computer-based testing washback, how can it be evaluated and how can this support practitioner research?", *Journal of Further and Higher Education*, Vol. 43/9, https://doi.org/10.1080/0309877X.2018.1471127. [85]

Kane, M. and R. Mislevy (2017), "Validating score interpretations based on response processes for the next generation of assessments", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*, Routledge. [5]

Kespaik, S., R. Dos Santos and F. Salles (2021), "Preparing and analysing log and process data in large-scale assessments", in Khorramdel, L., M. von Davier and K. Yamamoto (eds.), *Innovative Computer-based International Large-Scale Assessments – Foundations, Methodologies, and Quality Assurance Procedures*, Springer. [44]

Knox, J., B. Williamson and S. Bayne (2020), "Machine behaviourism: future visions of 'learnification' and 'datafication' across humans and digital technologies", *Learning, Media and Technology*, Vol. 45/1, https://doi.org/10.1080/17439884.2019.1623251. [86]

Kroehne, U., T. Deribo and F. Goldhammer (2020), "Rapid Guessing Rates across Administration Mode and Test Setting", *Psychological Test and Assessment Modeling*, Vol. 62/2. [68]

Kroehne, U. and F. Goldhammer (2018), "How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items", *Behaviormetrika*, Vol. 45/2, https://doi.org/10.1007/s41237-018-0063-y. [12]

Lee, Y. and S. Haberman (2015), "Investigating Test-Taking Behaviors Using Timing and Process Data", *International Journal of Testing*, Vol. 16/3, pp. 240-267, https://doi.org/10.1080/15305058.2015.1085385. [81]

Lee, Y. and Y. Jia (2014), "Using response time to investigate students' test-taking behaviors in a NAEP computer-based study", *Large-Scale Assessments in Education*, Vol. 2/1, https://doi.org/10.1186/s40536-014-0008-1. [67]

Lennon, R. (1956), "Assumptions Underlying the Use of Content Validity", *Educational and Psychological Measurement*, Vol. 16/3, https://doi.org/10.1177/001316445601600303. [47]

Li, H., C. Hunter and J. Bialo (2021), "A Revisit of Zumbo's Third Generation DIF: How Are We Doing in Language Testing?", *Language Assessment Quarterly*, Vol. 19/1, pp. 27-53, https://doi.org/10.1080/15434303.2021.1963253. [56]

Lindner, M. et al. (2017), "Identifying processes underlying the multimedia effect in testing: An eye-movement analysis", *Learning and Instruction*, Vol. 47, https://doi.org/10.1016/j.learninstruc.2016.10.007. [30]

Lindner, M. and S. Greiff (2021), "Call for Papers: "Process Data in Computer-Based Assessment: Opening the Black Box"", *European Journal of Psychological Assessment*, Vol. 37/3, https://doi.org/10.1027/1015-5759/a000658. [42]

Lindner, M. et al. (2018), "How Representational Pictures Enhance Students' Performance and Test-Taking Pleasure in Low-Stakes Assessment", *European Journal of Psychological Assessment*, Vol. 34/6, https://doi.org/10.1027/1015-5759/a000351. [7]

Li, Z., J. Banerjee and B. Zumbo (2017), "Response Time Data as Validity Evidence: Has It Lived Up To Its Promise and, If Not, What Would It Take to Do So", https://doi.org/10.1007/978-3-319-56129-5_9. [74]

Lundgren, E. and H. Eklöf (2020), "Within-item response processes as indicators of test-taking effort and motivation", *Educational Research and Evaluation*, Vol. 26/5-6, https://doi.org/10.1080/13803611.2021.1963940. [60]

Maddox, B. (2018), "Interviewer-respondent interaction and rapport in PIAAC", *Quality Assurance in Education*, Vol. 26/2, https://doi.org/10.1108/QAE-05-2017-0022. [28]

Maddox, B. (2017), "Talk and Gesture as Process Data", *Measurement*, Vol. 15/3-4, https://doi.org/10.1080/15366367.2017.1392821. [27]

Maddox, B. (2015), "The neglected situation: assessment performance and interaction in context", *Assessment in Education: Principles, Policy and Practice*, Vol. 22/4, https://doi.org/10.1080/0969594X.2015.1026246. [20]

Maddox, B. (2014), "Globalising assessment: an ethnography of literacy assessment, camels and fast food in the Mongolian Gobi", *Comparative Education*, Vol. 50/4, https://doi.org/10.1080/03050068.2013.871440. [26]

Maddox, B. et al. (2018), "Observing response processes with eye tracking in international large-scale assessments: evidence from the OECD PIAAC assessment", *European Journal of Psychology of Education*, Vol. 33/3, https://doi.org/10.1007/s10212-018-0380-2. [31]

Maddox, B., F. Keslair and P. Jayrh (2019), "Investigating Testing Situations", in Maddox, B. (ed.), *International Large-Scale Assessments in Education*, Bloomsbury. [29]

Maddox, B. and B. Zumbo (2017), "Observing Testing Situations: Validation as Jazz", https://doi.org/10.1007/978-3-319-56129-5_10. [21]

Maddox, B. et al. (2015), "An Anthropologist Among the Psychometricians: Assessment Events, Ethnography, and Differential Item Functioning in the Mongolian Gobi", *International Journal of Testing*, Vol. 15/4, pp. 291-309, https://doi.org/10.1080/15305058.2015.1017103. [59]

McNamara, T. and C. Roever (2006), *Language Testing: The social dimension*, Blackwell, Malden USA and Oxford. [19]

Messick, S. (1995), "Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning", *American Psychologist*, Vol. 50/9, https://doi.org/10.1037/0003-066X.50.9.741. [87]

Messick, S. (1989), "Validity", in Linn, R. (ed.), *Educational Measurement*, Macmillan Publishing Co. [48]

Michaelides, M., M. Ivanova and C. Nicolaou (2020), "The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items", *International Journal of Testing*, Vol. 20/3, https://doi.org/10.1080/15305058.2019.1706529. [39]

Mislevy, R. (2018), *Sociocognitive foundations of educational measurement*, Routledge. [80]

Mislevy, R. et al. (2014), *Psychometric considerations in game-based assessment*, GlassLab Research, Institute of Play. [70]

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. [14]

Oliveri, M., R. Lawless and R. Mislevy (2019), "Using Evidence-Centered Design to Support the Development of Culturally and Linguistically Sensitive Collaborative Problem-Solving Assessments", *International Journal of Testing*, Vol. 19/3, https://doi.org/10.1080/15305058.2018.1543308. [15]

Oranje, A. et al. (2017), "Collecting and Analyzing and Interpreting response times, eye tracking and log data", in Ercikan, K. and J. Pellegrino (eds.), *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*, Routledge. [23]

Padilla, J. and I. Benitez (2017), "A rationale for and demonstration of the use of DIF and mixed methods", in Zumbo, B. and A. Hubley (eds.), *Understanding and investigating response processes in validation research*, Springer Cham. [25]

Padilla, J. and J. Leighton (2017), "Cognitive Interviewing and Think Aloud Methods", https://doi.org/10.1007/978-3-319-56129-5_12. [53]

Pepper, D. et al. (2018), "Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics", *International Journal of Research and Method in Education*, Vol. 41/1, https://doi.org/10.1080/1743727X.2016.1238891. [24]

Piattoeva, N. and N. Vasileva (2021), "Infrastructuring the nation. Examining the role of national large-scale assessments in Russia", in Tröhler, D., N. Piattoeva and F. Pinar (eds.), *World Yearbook of Education 2022: Education, Schooling and the Global Universalization of Nationalism*, Routledge. [45]

Ranger, J., J. Kuhn and S. Pohl (2021), "Effects of Motivation on the Accuracy and Speed of Responding in Tests: The Speed-Accuracy Tradeoff Revisited", *Measurement*, Vol. 19/1, https://doi.org/10.1080/15366367.2020.1750934. [63]

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. [38]

Rojas, M. et al. (2021), "Assessing collaborative problem-solving skills among elementary school students", *Computers and Education*, Vol. 175, https://doi.org/10.1016/j.compedu.2021.104313. [72]

Salles, F., R. Dos Santos and S. Keskpaik (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. [9]

Sellar, S. et al. (2019), "Student preparation for large-scale assessments: A comparative analysis", in Maddox, B. (ed.), *International Large-Scale Assessments in Education*, Bloomsbury. [84]

Southgate, E. (2021), "Artificial intelligence and machine learning: A practical and ethical guide for teachers", in *Digital Disruption in Teaching and Testing: Assessments, Big Data, and the Transformation of Schooling*. [83]

Stoeffler, K. et al. (2020), "Gamified performance assessment of collaborative problem solving skills", *Computers in Human Behavior*, Vol. 104, https://doi.org/10.1016/j.chb.2019.05.033. [73]

von Davier, A. et al. (2017), "Interdisciplinary research agenda in support of assessment of collaborative problem solving: lessons learned from developing a Collaborative Science Assessment Prototype", *Computers in Human Behavior*, Vol. 76, https://doi.org/10.1016/j.chb.2017.04.059. [71]

Wise, S. (2020), "Six insights regarding test-taking disengagement", *Educational Research and Evaluation*, Vol. 26/5-6, https://doi.org/10.1080/13803611.2021.1963942. [4]

Wise, S. (2020), "The Impact of Test-Taking Disengagement on Item Content Representation", *Applied Measurement in Education*, Vol. 33/2, pp. 83-94, https://doi.org/10.1080/08957347.2020.1732386. [69]

Wise, S. (2019), "An Information-Based Approach to Identifying Rapid-Guessing Thresholds", *Applied Measurement in Education*, Vol. 32/4, https://doi.org/10.1080/08957347.2019.1660350. [66]

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. [37]

Wise, S. and X. Kong (2005), *Response time effort: A new measure of examinee motivation in computer-based tests*, https://doi.org/10.1207/s15324818ame1802_2. [36]

Wise, S., M. Kuhfeld and J. Soland (2019), "The Effects of Effort Monitoring With Proctor Notification on Test-Taking Engagement, Test Performance, and Validity", *Applied Measurement in Education*, Vol. 32/2, https://doi.org/10.1080/08957347.2019.1577248. [58]

Wyatt-Smith, C., B. Lingard and E. Heck (2021), *Digital disruption in teaching and testing: Assessments, big data, and the transformation of schooling*, https://doi.org/10.4324/9781003045793. [78]

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. [57]

Yaneva, V. et al. (2021), "Using Eye-Tracking Data as Part of the Validity Argument for Multiple-Choice Questions: A Demonstration", *Journal of Educational Measurement*, Vol. 58/4, https://doi.org/10.1111/jedm.12304. [54]

Zumbo, B. (2015), *Consequences, Side-Effects, and the Ecology of Testing: Keys to Considering Assessment In Vivo*. [22]

Zumbo, B. et al. (2015), "A Methodology for Zumbo's Third Generation DIF Analyses and the Ecology of Item Responding", *Language Assessment Quarterly*, Vol. 12/1, pp. 136-151, https://doi.org/10.1080/15434303.2014.972559. [55]

Zumbo, B., B. Maddox and N. Care (2023), "Process and Product in Computer-Based Assessments: Clearing the Ground for a Holistic Validity Framework", *European Journal of Psychological Assessment*, p. in press. [61]