

REVIEW

Open Access



The use of process data in large-scale assessments: a literature review

Ella Anghel^{1,2*} , Lale Khorramdel^{1,2} and Matthias von Davier^{1,2}

*Correspondence:
anghel@bc.edu

¹ Department of Measurement, Evaluation, Statistics, and Assessment, Lynch School of Education and Human Development, Boston College, 140 Commonwealth Avenue, Chestnut Hill 336, Chestnut Hill, MA 02467, USA

² TIMSS and PIRLS International Study Center, Lynch School of Education and Human Development, Boston College, Chestnut Hill, MA, USA

Abstract

As the use of process data in large-scale educational assessments is becoming more common, it is clear that data on examinees' test-taking behaviors can illuminate their performance, and can have crucial ramifications concerning assessments' validity. A thorough review of the literature in the field may inform researchers and practitioners of common findings as well as existing gaps. This literature review used topic modeling to identify themes in 221 empirical studies using process data in large-scale assessments. We identified six recurring topics: response time models, response time-general, aberrant test-taking behavior, action sequences, complex problem-solving, and digital writing. We also discuss the prominent theories used by studies in each category. Based on these findings, we suggest directions for future research applying process data from large-scale assessments.

Keywords: Large-scale assessment, Process data, Scoping review, Topic modeling

Introduction

Automatically gathering data about examinees' behavior, often referred to as log file or process data, has become increasingly common in computerized large-scale assessments (LSAs). Such data include examinees' response time to the test or its items, what they clicked on, and typing sequences, among others. This information can help improve item design, determine examinees' engagement, and inform inferences about the construct (Oranje et al., 2017), primarily if the use of such data is based on an existing theory of the cognitive processes underlying people's approach to a test (Brückner & Pellegrino, 2017). That way, log file data can provide evidence of the validity of the assessment (AERA, APA, & NCME, 2014). Moreover, they can help understand and reframe achievement differences in light of different test-taking strategies (Pohl et al., 2021).

The purpose of the current study is to provide a comprehensive review of common topics in existing research on log file or process data in LSAs. Given the value of educational and psychological theories in interpreting process data (Goldhammer & Zehner, 2017) and the relative lack of use of such theories in studies reporting such data in other contexts (Banihashem et al., 2018), we also focused on commonly used theories in the studies we reviewed. To meet these goals, we applied topic modeling (Blei, 2012), a natural language approach for automatically detecting topics in textual data.

This study would hopefully contribute to the existing literature in multiple ways: (a) reveal gaps in the literature (e.g., underexplored assessment domains or methods) by examining what has been studied so far, (b) highlight useful theories that could be relevant for future studies, (c) inform test developers of different ways in which log file data can be used to enhance assessment design, validation, and interpretation, and (d) introduce topic modeling as a method for literature review to the field of educational assessment and discuss its potential benefits and risks when used for that purpose.

Process data in LSAs

Log file data have become central in LSAs. They are routinely collected and published by many international and smaller-scale LSA programs, including the Programme for International Student Assessment (PISA; OECD, 2015), the Trends in Mathematics and Science Study (TIMSS; Fishbein et al., 2021), and the National Assessment of Educational Progress (NAEP; Bergner & von Davier, 2019). In their recent review of the literature, Reis Costa and Leoncio Netto (2022) discuss some seminal work in the field and show how they use different process variables to inform item-, person-, and system-level analyses.

There is no consensus on the definition of process data. Reis Costa and Leoncio Netto (2022), for example, define process data as “any type of information (e.g., response actions or timing) recorded on a computer platform into electronic files.” Namely, this view of process data relies on the technical aspects of how the data were collected. In contrast, Provasnik (2021) focuses on the content or purpose of the data and defines process data as “empirical data that reflect the process of working on a test question.” Such data are not necessarily recorded on a computer and may be gathered via cognitive interviews, eye tracking, brain imaging, etc. According to Provasnik (2021), then, Reis Costa and Leoncio Netto’s (2022) view of process data fits better to the term “log file data”, which highlights the source of the data rather than their purpose.

This work focuses on process data captured automatically in log files, log file data according to Provasnik (2021). However, while we agree that the distinction between log file and process data is crucial, it is relatively new and has not been consistently applied in the literature yet. It seems like much of the existing literature refers to log file data when using the term process data either with (e.g., Bergner & von Davier, 2019; Goldhammer et al., 2021a, 2021b) or without acknowledging their difference (e.g., Chen et al., 2019; Goldhammer et al., 2017a, 2017b).

In order to reflect that, we use the term “process data” to refer to any data automatically collected about test-takers’ response process. Such data typically include timing data (e.g., response time per item, the time between clicks) and the actions performed by the test-taker (keystrokes, mouse clicks, selection, navigation, etc.). They do not include automatically-collected data that goes beyond the response process like the device and setting in which the assessment took place (i.e., access-related data; Kroehne & Goldhammer, 2018), metadata, or the responses themselves. While such data are interesting and valuable, they are beyond the scope of our definition.

Theoretical models for process data in assessments

The underuse of theories is often discussed as an issue in learning analytics research (Khalil et al., 2023). Perhaps due to the wealth of data or the rapid methodological advancements of recent years, many studies using process data or other clickstream data seem to be purely empirical, focusing on methods and findings rather than on testing theoretical predictions. While data-driven studies are certainly important for describing phenomena and can be useful for informing theory development, particularly in nascent fields like learning analytics, theories are key for generating hypotheses, interpreting findings, and applying them in practice (Banihashem et al., 2019).

Theory seems to play a more important role in the field of educational assessment. For example, the most commonly acceptable definition of validity (AERA, APA, & NCME, 2014) refers to both empirical evidence and theory as important aspects of validity. Theoretical models describing how the target construct is structured, developed, and manifested are key in assessment construction (Rupp et al., 2010), and inform task design, validation, and interpretation (Borsboom et al., 2004).

Kane and Mislevy (2017) show how theories and process data can be used together for validation purposes. Cognitive theories that describe the mechanisms behind how performance in a certain task (e.g., responding to an item) is produced can be used to guide process data interpretation. Process data can then be used as evidence for said mechanisms, or as evidence that the mechanisms are not being used in practice (e.g., students are guessing). Motivational theories can also be used to explain why unexpected strategies are used, potentially contributing to preventing such issues in the future.

In spite of their importance in assessment, it is unclear whether theories are being used in practice in studies involving process data. As process data has not been consistently collected and studied until relatively recently, theories are likely underused in studies using process data, similar to the status of the learning analytics field as a whole. Therefore, one of the purposes of this study is to examine whether and which theories have been used in the field so far.

Topic modeling for literature reviews

Given the multitude of studies applying process data in the context of LSAs, we decided to apply a natural language processing (NLP) approach to classify the studies. Specifically, we used topic modeling (Blei, 2012), an unsupervised learning method used to detect latent commonalities, or topics, in a large group of documents (in our case, research papers). This approach seems to be particularly useful for a review that focuses on what the main topics studied in a field are, and it has been applied in contexts similar to ours. For instance, Chen et al. (2022) used topic modeling to identify themes in learning analytics research and found that these papers covered topics like web services for learning, flipped learning, and assessment.

Topic modeling is a name for a group of methods (see Vayansky & Kumar, 2020 for a review). We chose to use structural topic modeling (STM; Roberts et al., 2016), which allows for the inclusion of metadata. This approach was preferred because it enabled us to include the articles' first author as an additional data source when identifying topics. Since authors tend to write about similar topics, we believe including the first author as

an additional information source could lead to more stable topics. We now turn to an intuitive introduction to topic models. A more rigorous treatment is available elsewhere (e.g., Blei, 2012).

In any topic model, it is assumed that there is a known number of topics in the data, and each topic is related to words with different probabilities. So, in a dataset of restaurant menus, there can be a topic such as “breakfast” which is related to the occurrence of words like bagel or eggs, and a topic such as “drinks” which is indicated by words like tea or soda on the menu. Then, each document (in our case, menus) contains a number of words and the distribution of these terms emerges from a mixture of topics with different probabilities. So, a menu of a bagel shop is more likely to be generated from the topic “breakfast” than from the topic “drinks”, and it is more likely to include the words bagel or eggs. A menu of a café might be a mixture of both categories and may include eggs but also tea. In structural topic modeling, information about the restaurant’s location or head chef could also inform the identification of topics. For instance, a menu of a restaurant in New England might be more likely to be of the topic “seafood” than “barbecue.” In reality, the probabilities that words and documents are related to certain topics are unknown and have to be estimated from the observed documents.

In our case, the model identifies groups of articles that share similar content (i.e., words) and uses these common words to derive the topics and their relationships. Then, each paper is assigned a set of probabilities to be representative of these topics. Some papers may clearly belong to one single topic and not to the others (as was the case for our bagel shop menu), so they will be assigned with a high probability of being in that topic and a low probability of being in the others. Other papers may represent a mixture of several topics. Regardless, topic modeling is commonly used for grouping large collections of text documents by their content, making it particularly appropriate for literature reviews. In this study, we used topic modeling to identify common themes in the extant research on process data in LSAs.

The current study

There are no comprehensive literature reviews of the use of process data in LSAs. Existing reviews focus on the use of technology in LSAs (Zenisky & Sireci, 2002) or the use of process data in related fields (e.g., higher education; Viberg et al., 2018). In their recent work, Reis Costa and Leoncio Netto (2022) review selected papers related to process data in international LSAs. While valuable, this work does not introduce the full picture of how process data is used in LSAs research, as smaller-scale assessments are some of the most advanced users and researchers of such data (e.g., Bergner & von Davier, 2019). In addition, Reis Costa and Leoncio Netto (2022) use an ecological framework to analyze their target studies and do not examine the theories used in the studies themselves.

In the current review, we explore common topics in the literature on process data in LSAs as a whole. We also looked at whether and which theories have been used in existing studies. We believe that this is the first step in understanding what is already known, what remains to be explored, and how theories can inform future studies, thus contributing to other researchers interested in this topic. So, our research questions are:

RQ1: What are the main topics in studies involving process data in LSAs?

RQ2: To what extent are they informed by psychological or educational theories?

Methods

Our review follows Arksey and O'Malley's (2005) framework for conducting rigorous scoping reviews. Scoping reviews aim to provide a snapshot of the field and describe the range of relevant studies without evaluating their quality. Since we were interested in the various topics being explored in the literature and did not aim to comment on their quality, and given the diverse foci of the relevant studies, we considered this approach to be appropriate. Arksey and O'Malley's framework has five stages, which were repeated iteratively until a final sample of studies was identified. The five stages are presented below.

Identifying the research question

In this stage, the authors define their research questions and all of their variables. Our research questions are presented above. Our main variable, process data, is defined as any data automatically collected about test-takers' response process. As discussed above, while we acknowledge the complexity and disagreement in the literature on what constitutes process or log file data, we found this definition to be simple and inclusive of the studies we were interested in.

Definitions of LSAs can also be complex (see a discussion in Khorramdel et al., 2023). However, we again chose a rather simple definition of LSAs which allowed us to include a wide variety of assessments and therefore, studies: LSAs are any achievement or knowledge assessments administered to many test-takers, at least to an entire school district. Future studies might focus on more specific LSA definitions, e.g., those focusing on monitoring educational systems.

We posed no limitations on the age of the participants. We included both children (e.g., school achievement tests like TIMSS and PISA) and adults (e.g., licensure tests or the Program for the International Assessment of Adult Competencies [PIAAC]). Since we were not interested in the scale of the study but rather the type of assessment used, we included studies that used publicly available LSA items but were small-scale themselves (e.g., fewer than 100 participants).

Finally, we wanted to focus on educational assessments, so we chose to focus our search on the content domains of reading literacy, mathematics and numeracy, and science. These domains were selected as they are often the focus of LSA, though other content domains could also be examined in the future. We included studies not directly measuring these domains only if they were administered together with and were related to a test of one of these domains (e.g., a survey on the effort the student put into a mathematics assessment).

Identifying relevant studies

In this stage, the strategies used to find studies for review are defined. Our main method was to search electronic databases. We searched Scopus, ERIC, and Google Scholar from May to June 2022. We included several groups of keywords separated by the Boolean "AND". The first involved the type of data we were interested in, and included the following terms separated by "OR": "process data", "response process", "timing", "clickstream", "keystroke", "log data", "log files", "activity logs", "trace data", "event data", "response time",

“speededness”, “telemetry”, “sequence”, “learning analytics”, and “data mining”. The second group was related to assessment and included the terms “assessment”, “testing”, and “measurement” (separated by “OR”). Next, as we wanted to focus on specific assessment domains, we added these terms separated by “OR”: “reading”, “language”, “literacy”, “writing”, “numeracy”, “math”, “quantitative”, “science”, and “stem”. Finally, given a large number of irrelevant studies that were found, we added the context we were interested in, “education” OR “school”. In platforms that allowed for more focused searches, these terms were searched only in the title, keywords, and abstract of the papers to make sure only relevant results were identified. Otherwise, the full manuscript was searched. These steps resulted in 2718 articles from Scopus, 1094 from ERIC, and 1107 from Google Scholar, but after removing duplicates, 4677 papers remained.

After some steps related to study selection (see below), we examined the references of the remaining studies for titles that included the relevant keywords. We read those based on the study selection criteria and, if they were relevant, searched their references as well. This procedure was repeated until no new studies were added. This method added 158 papers to our list.

Study selection

Before the initial search, we developed several criteria to narrow down the results. First, we only included studies relevant to our RQ, namely, that relied on at least one type of process data in the context of an educational LSA. Note that while our initial content domains and search terms only included reading, math, and science, we identified many studies involving complex problem-solving (CPS) assessments, a topic related but not directly a part of any of these domains. We decided to include these studies given their ubiquity in the literature and CPS’ importance in the content areas we focused on (OECD, 2014).

We also limited our search to English-language papers that had some empirical components. That is, we included review papers or model development papers only if they included an application on empirical data. We did not limit ourselves in terms of the publication platform, except we did not include full books, dissertations, and test technical reports due to their volume; chapters in books and reports of studies done by test developers were included. Based on these criteria we read the papers’ titles and abstracts to determine their relevance. We excluded 4603 sources (see details in Table 1) and

Table 1 Details on the excluded sources

Reason for exclusion	Number of sources
Beyond the scope (not LSA or not process data)	4306
Not empirical	123
Conference proceedings	96
Books, dissertations, or technical reports	55
Preprints or work in progress	11
Not in English	9
Unavailable to the authors	3

Some studies could be excluded due to multiple reasons, but we only recorded one reason per source

retained 74. Together with the 158 identified from these studies' references, our final sample included 232 articles.

Charting the data

In this stage, the articles are coded for the variables of interest, usually by carefully reading them and recording variables relevant to the research questions. In our case, we first performed the topic modeling as a way of collating the articles (see “[Collating, summarizing, and reporting the results](#)” Section). We then manually coded their publication year and venue, first author, and the assessment they used. We also wrote a brief description of their methodology and some key findings. As we were interested in the articles' theoretical basis, we also documented their use of theories. All of these were used to summarize the studies within each topic, as described in “[Collating, summarizing, and reporting the results](#)” Section.

Collating, summarizing, and reporting the results

As described in “[Topic modeling for literature reviews](#)” Section, we applied a structural topic model for an initial summarization of the data. We used the *stm* R package (Roberts et al., 2019) to extract the topics. Note that the topic modeling was conducted on the papers' abstracts and not the full papers. This was done in order to focus on the most important terms within each paper rather than extraneous information that might result in meaningless topics (e.g., topics focusing on a specific assessment as opposed to the study's content). We also excluded executive summaries which are usually longer than abstracts. So, we excluded 11 papers that did not have an abstract, and only 221 papers were included in the analysis.

We then performed several data-cleaning steps. We removed citations and section headings within structured abstracts (e.g., “background”, “conclusions”). This was done to minimize common phrases among papers that are not similar in their content. We also expanded contractions and spelled out all acronyms with the exception of test names (e.g., we kept TOEFL instead of Test of English as a Foreign Language, PIRLS for Progress in International Reading Literacy Study, etc.), as these are known by their abbreviated name, and expanding it could result in a wrong classification. Next, we used the automatic *stm* cleaning functions which perform standard NLP pre-processing such as setting all words to lowercase, removing punctuation and stopwords (function words like “the”), and stemming the words (recoding them to their most basic form, e.g., “responding” turned to “respons”).

Following these steps, we had to choose the optimal number of topics for the analysis. We did so by estimating models with five to ten topics. These numbers were selected because they would, assuming a uniform distribution, lead to an expected number of 20–40 papers per topic, which we considered reasonable for succinctly summarizing the literature. These five to ten topic models were then evaluated using their semantic coherence and FREX (frequency-exclusivity), two commonly used measures of topic models' quality (e.g., Pandur et al., 2020). Semantic coherence represents how often the most common words of a topic occur with other common words, or how consistently these words cooccur. In a good-quality model, we expect

to see high levels of coherence. For example, if a topic's most common words are "log" and "file", we would like them to appear together in most of the papers assigned to this topic.

A topic's exclusivity is the extent to which the most common words in a topic are distinct from the words in other topics. This is a measure of how unique the topics are, and we are interested in exclusive, easily distinguishable topics. For example, if one topic had the words "response" and "time", we would like these words to be rare in other topics. Exclusivity alone, however, is inflated by rare words; if a word appears only in two documents in the whole dataset and the documents belong to the same topic, that topic will have an unjustifiably high exclusivity value. So, we used FREX which takes into account the words' frequency in the data overall when determining how different topics are. However, it is important to note that some overlap between topics should be expected even when FREX values are high, especially in this case where the texts share many similar words. A more formal treatment of semantic coherence and FREX is available in Mimno et al. (2011) and Bischof and Airolidi (2012), respectively.

Note that while we want models with high topic coherence as well as high topic FREX, there is usually a tradeoff between the two. Models with few topics tend to have high coherence values but have low FREX values because, while it is likely that the common words in each topic appear in many documents within that topic, it is also likely that they are not unique to that topic. There are also no acceptable thresholds for these measures as they heavily depend on the specific dataset. So, to select the optimal model, we plotted the coherence and FREX values for each of the five to ten topics models we estimated and chose a model that had relatively high values on both measures. We also preferred a more parsimonious solution, namely, with fewer topics.

Once we selected the number of topics we estimated our final model. This resulted in a vector of probabilities for each paper representing how likely it is to belong to each of the topics. For each paper, we chose the topic with the highest probability and assigned the paper to that topic. Then, we interpreted the models based on their most common words and some representative examples. The resulting topics and the content of the papers assigned to each topic are reported in the Results section. After each paper was assigned to a topic, we read all the abstracts in each topic to identify prominent subtopics. This way, the topic model assisted us in the initial identification of themes in the reviewed studies but we were also able to delve into more detailed types of studies. We also searched specifically for theories and findings in the respective sections of the papers. Table 2 presents this procedure.

Table 2 Summary of the data summarization stage

Stage	Method	Data source
Identifying topics	Topic modeling	Abstracts
Identifying subtopics	Human coding	Abstracts
Identifying theories, methodologies, and key findings	Human coding	Full papers

Results

Figure 1 presents the mean semantic coherence and FREX of the models with five to ten topics. Based on these results and an examination of the topics themselves, we chose a model with six topics. The topics' most common words and representative examples led us to label the topics as response time models, response time-general, aberrant test-taking behavior, action sequences, complex problem-solving, and digital writing.

Table 3 presents the topics' most common words, first authors, publication venues, and assessments used. The most prolific first authors in our sample overall were S. L. Wise (with 19 papers), W. J. van der Linden (8), and F. Goldhammer (7). The most common publication venues were the *British Journal of Mathematical and Statistical Psychology* (14), *Applied Measurement in Education* (13), and *Frontiers in Psychology* (12). The most common assessments used were PISA (40), PIAAC (26), and Measures of Academic Progress (MAP) Growth (20), though 32 studies did not name the specific assessment used. Table 4 presents descriptive statistics of the probabilities of the papers assigned to each topic. Appendix A presents the papers' probabilities of being in each topic and the topic they were assigned to. Below we present these topics and the findings of studies assigned to these topics.

Response time models

This topic includes mostly methodological papers where a new model for response time analysis is developed or evaluated. They were published between 1997 (Jansen, 1997) and 2021 (DeCarlo, 2021; Rios & Soland, 2021). These papers are rarely based on a theoretical foundation but many of them expand on existing measurement models (see below).

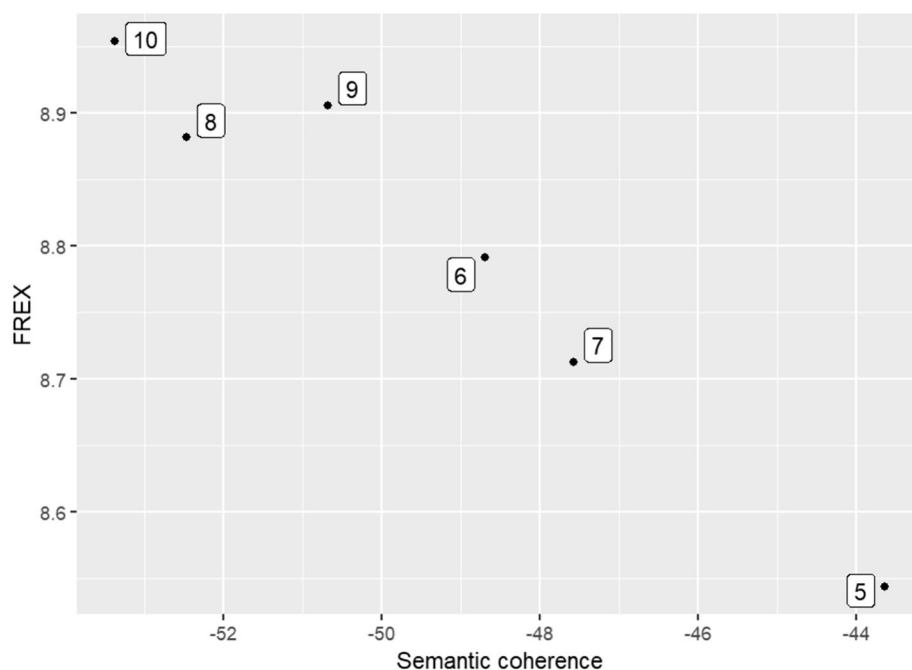


Fig. 1 The topic models' coherence and FREX

Table 3 The six topics and common unique words, first authors, publication venues, and assessments

Topic (%)	Top-5 stems by FREX	Top-3 first authors (frequency)	Top-3 publication venues (frequency)	Most common named assessment (frequency)
Response time models (19%)	Paramet, model, mixtur, hierarchy, fit	van der Linden, W. J. (8), Bolsinova, M. (5), Wang, C. (5)	British journal of mathematical and statistical psychology (12), Psychometrika (9), Journal of educational and behavioral statistics (8)	PISA (6)
Response time-general (14%)	Latenc, choic, multipl, answer, drag	-	The annual meeting of AERA (3)	Graduate record examination (GRE; 3), PISA (3)
Aberrant test-taking behaviors (22%)	Disengag, guess, rapid, effort, take	Wise, S. L. (15), Rios, J.A. (5), Soland, J. (5)	Applied measurement in education (9), Large-scale assessment in education (5), Educational and psychological measurement (4), International journal of testing (4)	MAP (17)
Action sequences (16%)	Action, sequenc, technolog, data, mine	He, Q. (6)	Frontiers in psychology (4)	PIAAC (15)
Complex problem-solving (13%)	Navig, complex, solv, strategi, problem	Goldhammer, F. (4), Eichmann, B. (3), Greiff, S. (3), Naumann, J. (3)	Computers in human behavior (5), Journal of educational psychology (5), Frontiers in psychology (4)	PISA (16)
Digital writing (16%)	Keystrok, essay, revis, paus, write	Deane, P. (4), Zhang, M. (4), Bennett, R. E. (3)	ETS research report series (9), Studies in second language acquisition (3)	Cognitively based assessment of, for, and as learning (CBAL)

We only present authors, publication venues, and assessments that had more than two instances. In cases of ties with more than two occurrences, we presented all tied authors, publication venues, and assessments in Alphabetical order

Table 4 Topic probabilities' descriptive statistics

Topic	Papers assigned to the topic				All papers ($n = 221$)		
	<i>n</i>	Mean	Standard deviation	Range	Mean	Standard deviation	Range
Response time models	42	.85	.15	.52-.99	.19	.33	.00-.99
Response time-general	31	.71	.22	.33-.98	.15	.26	.00-.98
Aberrant test-taking behaviors	48	.78	.20	.26-.98	.20	.33	.00-.98
Action sequences	36	.77	.19	.26-.98	.17	.29	.00-.98
Complex problem-solving	28	.80	.19	.42-.99	.14	.27	.00-.99
Digital writing	36	.83	.17	.43-.99	.15	.31	.00-.99

n is the number of papers where this topic had the maximum probability

Many of the articles in this group introduce new models combining item responses and response times. They are usually a type of item response theory (IRT) model (e.g., Jeon & De Boeck, 2019; Pokropek, 2016; Wang & Hanson, 2005) or mixture models that identify different populations within the data (e.g., Liu et al., 2020; Ulitzsch et al., 2020c; Wang & Xu, 2015). Of specific note given its impact on the field is van der Linden's (2007) hierarchical speed-accuracy framework. This framework suggests combining any IRT model that suits the data, a model describing the response time distribution (e.g., a lognormal distribution; van der Linden, 2006), and a higher-level model that accounts for any dependencies among the parameters of these models at the population level (e.g., high-ability test-takers also tend to work faster).

Many of the rest of the articles extend or develop tools to be used within van der Linden's (2007) framework. Some relax van der Linden's model's assumptions, specifically that response time and accuracy are conditionally independent, allowing for correlations between response time and accuracy even after accounting for the overall ability and speed traits (Bolsinova et al., 2017a, 2017b; Meng et al., 2015). Others extend it to a multidimensional case (Lu et al., 2020; Man et al., 2019; Zhan et al., 2018). And yet others propose other additions such as fit statistics (Fox & Mariani, 2017), hypothesis testing (van der Linden & Glas, 2010), or invariance testing procedures (Glas & van der Linden, 2010) that can be applied to that model.

Response time-general

This group included papers that used response time not necessarily as a part of a methodological paper, though other than the use of response time, we were not able to find another common theme within this group. These papers were published between 1994 (Bergstrom et al., 1994; Parshall et al., 1994) and 2022 (Araneda et al., 2022). Most of these studies did not rely on theoretical grounds, but some mention the dual-processing model (Evans & Stanovich, 2013). This model divides higher-order thinking into processes that are performed rapidly and intuitively or slowly and methodically.

A common subtopic links response times with item or person characteristics. Some item characteristics that were found to be associated with longer response times are item difficulty, discrimination, length, late position in the test, and cognitive complexity (Bergstrom et al., 1994; Halkitis et al., 1996; Zenisky & Baldwin, 2006). Findings are less consistent with respect to person characteristics (Parshall et al., 1994; Zenisky &

Baldwin, 2006). Finally, studies had different conclusions regarding the link between ability or response accuracy and response times, where some found that longer response times are associated with lower or higher ability whereas others found no association (Bergstrom et al., 1994; Bridgeman & Cline, 2000). This might be attributed to the behaviors described in the dual processing theory, where the association between ability and speed depends on whether the task requires slow vs. fast processing (also known as the time-on-task effect; Goldhammer et al., 2014).

Another set of studies involved item design. These studies looked at innovative item designs or layouts and how they differ from traditional multiple-choice items, mainly in terms of response time. For example, they tested the effects of innovative designs on response accuracy and speed with different results depending on the item type (Ponce et al., 2021; Saß et al., 2012). Other studies compared the information and efficiency (information per time unit) of these items, finding that they are at least as informative as multiple-choice items (Jodoin, 2003; Ponce et al., 2020; Wan & Henly, 2012).

Finally, a set of studies focused on measurement issues related to response time. These studies, for example, determined the reliability of a speed factor (Davison et al., 2012) and its invariance (Shin et al., 2020). They also used different methods to detect aberrant behaviors in a test (speededness or item preknowledge) and their potential impact on test design (Margolis et al., 2020; Qian et al., 2016).

Aberrant test-taking behaviors

The topic of Aberrant test-taking behaviors includes papers identifying and/or using information about participants' actions that reflect disengagement or guessing. Unlike a subgroup of the previous topic, these studies did not necessarily focus on response time to identify aberrant test-taking behaviors (though many of them did). The studies assigned to this group were published between 1997 (Schnipke & Pashley, 1997; Schnipke & Scrams, 1997) and 2021 (e.g., Gorgun & Bulut, 2021).

About half of the studies do not mention theories. Among the rest, some prominent theories are the expectancy-value theory (Wigfield & Eccles, 2000) and the demands-capacity model (Wise & Smith, 2011). Broadly, the expectancy-value theory suggests that whether a person engages in a behavior (i.e., working hard on a test) depends on the behavior's subjective value to that person and their expectations that they can succeed in performing it. The demands-capacity model posits that the effort an examinee exerts when responding to an item depends on what the item demands or the resources it requires and on the examinee's capacity or willingness to work on the item. It is easy to see how both theoretical frameworks are relevant when studying examinees' test-taking efforts. Schnipke and Scrams' (1997) conceptualization of test-taking behavior as either solution behavior or rapid guessing behavior is also often mentioned in this category.

A common subtopic in this group involves the comparison of different response time thresholds for identifying non-effort in a given item or examinee; usually, non-effort is defined as spending less than a certain amount of time on an item or a group of items, and there are many methods in determining that amount of time. The most common item thresholds were absolute (e.g., below two seconds; Soland et al., 2019a, 2019b) or normative (e.g., 10% of the average time all test-takers spend on the item; Wise & Ma, 2012). There were other, more complex ones, for example, a visual inspection of

the empirical response time distribution to identify item response times that were more common than expected when assuming solution behaviors (Lee & Jia, 2014), an examination of when response accuracy is below chance (Wise, 2019), fitting mixture models that separate effortful from non-effortful responses (Soland et al., 2021), and combining response times with the number of actions a test-taker took (Sahin & Colvin, 2020). Each of these methods has its advantages and disadvantages, and different studies disagree on which one is the best and in what context.

Methods for identifying disengaged test-takers are less involved and are usually based on the percent of items where disengagement was identified based on one or more of the aforementioned thresholds (e.g., a person who was determined to be non-effortful in over X% of the items is considered non-effortful, where X is often 10 or 15; Wise, 2014). A different approach which is more suitable for survey items was proposed by Soland et al. (2019a, 2019b), where multiple factors like responses to reverse-coded items and distribution of the selected response options were examined to determine participants' effort.

Another subtopic involved work on using statistical methods for identifying disengagement. For instance, Ulitzsch et al. (2020a) developed a model for identifying disengaged participants based on rapid guesses and non-responses, and Wise and Kuhfeld (2021a) developed a method for identifying semi-effortful responses. Related, some works compared different methods or models for identifying disengagement in terms of their convergent and discriminant validity (Rios et al., 2014) or fit to data (Wise & Kingsbury, 2016). Other studies explored the statistical effects of non-effort on ability estimation at the individual- (Rios et al., 2017) or school level (Wise et al., 2020) or in comparing group scores (Rios, 2021) or students' growth (Wise, 2015).

Other papers were more applied rather than model- or method-focused. For instance, several studies compared the levels of disengagement or rapid guessing across groups, including by country, language, school track, gender, and race (Goldhammer et al., 2016; Kroehne et al., 2020; Lindner et al., 2019; Los et al., 2020; Soland, 2018a). Other studies tested the link between disengagement and performance (e.g., Pools & Monseur, 2021) or other variables of interest, including student characteristics (e.g., demographics, school achievement, and school engagement; Jensen et al., 2018; Soland & Kuhfeld, 2019; Soland et al., 2019a, 2019b) and item properties (e.g., item difficulty, length, position, or content; Kroehne et al., 2020; Michaelides et al., 2020; Setzer et al., 2013). Note that these studies are different from those mentioned above as they do not test the effects of disengagement on ability estimates but rather use the disengagement measure itself and compare it across groups or use it for prediction purposes.

Action sequences

This topic includes studies that used various methods to examine test-takers' action sequences, namely, the number and order of actions (e.g., clicks) within the assessment platform. For example, Lee and Haberman (2016) used these data to see if test-takers respond to items in the intended order. The studies were published between 2010 (Bennett et al., 2010) and 2021 (e.g., He et al., 2021). They rarely rely on a theory except for a few studies using models of the constructs they target.

The largest subtopic in this group involved score prediction based on process data. Some studies used distance-based methods, namely, they identified how different examinees' behavior was from the most efficient solution and linked it to performance (Hao et al., 2015; He et al., 2019a, 2019b). Others identified specific actions or series of actions that were associated with success in the item goal-directed behavior (He & von Davier, 2015), number of attempts (Stadler et al., 2019), systematic control of variables (known as the vary-one-thing-at-a-time or VOTAT strategy; Tschirgi, 1980; e.g., Han et al., 2019), etc.

Group comparison and test-takers clustering were also common types of studies. Some studies compared the actions taken by groups based on demographics (Bennett et al., 2010; Liao et al., 2019) or by their performance (Zehner et al., 2020). Others identified groups within the data. For instance, He et al., (2019a, 2019b) detected three groups based on actions in a CPS item and labeled them as low-, medium-, and high-effort, and Xu et al. (2018) found different incorrect strategies when clustering participants' action sequences.

Finally, there were several methodological studies included in this group. Some studies explored the use of dimension reduction techniques (Tang et al., 2020, 2021) before using the reduced data for other purposes such as prediction. Other studies demonstrated models (e.g., Cui et al., 2020 and the adaptive neuro-fuzzy inference system, Kroehne & Goldhammer, 2018 and finite state machine, or Shu et al., 2017 and Markov-IRT). And others compared the performance of several models (Qiao & Jiao, 2018; Ramalingam & Adams, 2018).

Complex problem-solving

The papers assigned to this topic are mostly about test-takers' interactions with items designed to measure CPS skills. Like the previous group, many of them used action sequences, but here the focus was on CPS as a construct and not the type of data, distinguishing this topic from the previous one. They were published between 2013 (Goldhammer et al., 2013; Greiff et al., 2013) and 2022 (Ulitzsch et al., 2022). Most of these studies did not rely on an existing psychological or educational theory and were more exploratory, but the dual-processing theory was mentioned more than once (e.g., Scherer et al., 2015). In addition, some papers sometimes mention frameworks of the target construct that were used to design the items or identify key response processes (e.g., the OECD framework for CPS as consisting of knowledge acquisition and application; Eichmann et al., 2019).

Most of the studies in this category used actions taken by examinees and linked them to their performance. Some of the studies focused on relatively simple predictors like the number of actions performed and response time (Scherer et al., 2015; Vörös & Rouet, 2016). Others looked at more complex patterns based on action sequences that had real-world interpretations. For example, Greiff et al. (2015) used the log data in a PISA CPS item to identify the VOTAT strategy. This strategy was then found to be related to individual- and country-level performance on that item. VOTAT was very commonly used in the reviewed studies and consistently found to predict better performance (e.g., Wüstenberg et al., 2014). Other problem-solving behaviors were detected, as well (e.g., preparation vs. execution; Eichmann et al., 2020a, 2020b).

Other studies used process data to detect latent groups or to compare the behaviors of existing groups based on performance or demographics. For instance, based on their action sequences in a CPS item, Teig et al. (2020) distinguished between those who consistently used VOTAT, those starting to use VOTAT, and disengaged test-takers. Eichmann et al. (2020a, 2020b) compared different groups' performance and strategies taken in a PISA CPS item. They found that boys interacted more with the item and had more exploration behaviors, which explained why boys were more successful than girls on that item.

Digital writing

In this category, the studies involved writing assessments and examinees' interactions with them. These studies were published between 2008 (Crawford et al., 2008) and 2022 (Bennett et al., 2022; Gong et al., 2022; Talebinamvar & Zarrabi, 2022). Most studies relied on some theoretical model of writing. One example is Hayes' (2012) work which describes the writing stages as planning, translating (the plan into language), transcribing (into written language), and evaluating (e.g., Guo et al., 2020). Another example is Kellogg's (1996, 2001) works where writing consists of formulation, execution, and monitoring (e.g., Révész et al., 2019). As many studies use data from a specific assessment, the CBAL, they also often describe the model used to design this assessment (Deane et al., 2015; e.g., Zhang et al., 2016).

Studies about digital writing covered several subtopics, the largest one involved comparing the writing patterns of existing groups. For example, Bennett et al. (2021) found gender differences in the response process (typing speed, productivity, and specific types of editing behaviors) even after controlling for performance. Deane, Roth, et al. (2018) took a different approach and manually classified test-takers' behaviors into stages (e.g., drafting, copy-editing) and found that they were associated with different typing behaviors. For instance, during drafting, test-takers wrote fluently but paused between sentences (reflecting thinking about the next sentence), but during copy-editing, typing was slow and edits were focused on individual words.

Other prominent subtopics were the association of process features with performance and identifying behavioral patterns based on those features. For instance, Bennett et al. (2020) found that high-performers typed quicker, started more words, and spent more time on the task, but did not differ from low-performers in terms of editing behaviors. In an example of finding behavior patterns, Talebinamvar and Zarrabi (2022) used typing pauses and typing fluency (among other measures) and identified five clusters of test-takers: strategic planners, rapid writers, emerging planners, average writers, and low performers. These clusters were associated with performance – strategic planners had the highest performance, and low performers had the lowest.

Discussion

This review explored common topics within the literature on process data in studies about LSAs, applying a topic modeling approach for the initial theme extraction process. We identified six overarching topics, response time models, response time-general, aberrant test-taking behavior, action sequences, CPS, and digital writing. We also discussed the use of theory in the reviewed topics as well as common subtopics within them.

Several observations can be made based on our findings. First, it seems that some of our topics focused on response time and some focused on other data sources. The ubiquitous use of response time is unsurprising since, unlike typing, for example, total item response time is relevant in any type of item response process, and is relatively easy to record. This is also evident by the dates of the studies, as the topics related to response time include much older studies in comparison with the other topics (e.g., the response time–general group has studies from 1994 while the oldest study in the digital writing group is from 2008). Notably, studies rarely use other types of response time such as the time before the first action, the time between revisions, etc. That information is more commonly used in papers on digital writing and might be useful in analyzing other types of items as well, including relatively simple multiple-choice items. So, we suggest that this information should be collected in any item and made available to researchers.

Second, many of the reviewed studies were methodology-focused, namely, they proposed, evaluated, or improved statistical models for estimating examinees' abilities. Again, perhaps unsurprisingly given their availability, most of these models augment existing measurement models with item response time data rather than other types of process data. Future studies could explore the use of other timing or action data to enhance ability estimates.

Third, we presented here topics that were common in the literature, but what is less studied could also be of interest to researchers. While we identified specific topics related to writing and CPS, reading and mathematics did not have their specific group, but were rather included in other topics (e.g., Salles et al.,s, 2020 study on a mathematics assessment was included in the Action sequences topic). This suggests that there are relatively few studies on these constructs or that they do not use a unique terminology distinguishing them from other studies. Assuming this is not an artifact of our search strategy, we can hypothesize a few reasons for this finding.

With respect to reading, it is more difficult to capture the process automatically without eye-tracking devices or other types of specialized hardware compared to typing or reacting within a given time. Clicks, mouse movements, and scrolling are more indirect measures, and while they can still be useful in studies on digital reading, they may be more obvious choices for capturing the writing process or online navigation. Nevertheless, even simple measures such as response time can serve to assess reading fluency as described in theories of reading such as the Information Processing Theory (LaBerge & Samuels, 1974), and should be used more often in reading LSAs.

That there was not a separate topic for mathematics was surprising because some mathematics subjects seem to lend themselves well to studies involving process data (e.g., geometry; von Davier et al., 2023), and can be easily related to elements of relevant theories such as the Mathematical Problem-Solving model (e.g., self-monitoring, specific problem-solving strategies; Schoenfeld, 2013). Other researchers may want to explore this subject further.

Finally, we found very few studies on the use of process data in surveys. Admittedly, these were not at the center of our exploration, so it is possible we did not include relevant studies in the current review. Nevertheless, we wanted to emphasize that process data have the potential to assess (dis)engagement in surveys in a way that

conforms with existing motivational theories and that can inform estimates of the targeted constructs (e.g., Soland et al., 2019a, 2019b).

Related to theories, the field of learning analytics is seeing a transition from data-driven to theory-driven research, as theories “contribute to understanding and explaining phenomena or make visible causal relationships between variables, as well as allowing us to make rational inferences about future events” (Khalil et al., 2023). We were interested in the extent to which similar studies on LSAs rely on educational or psychological theories. While many of the reviewed studies were data-driven or methodology-focused, lacking a theoretical basis, several other studies did use relevant theories, either motivational (expectancy-value theory) or cognitive (dual processing theory, demands-capacity model). Some studies that focused on validating a specific assessment also introduced theories on how its target construct should be measured. Given how educational assessments start with defining what one wants to measure, and process data is often used to gather evidence related to this question (Knight & Shum, 2017), it makes sense that the use of theories is relatively common in the field of educational measurement, but there is still room for improvement.

This study as a whole demonstrates the strengths of using topic modeling for literature reviews. Having themes automatically generated from abstracts saved time on multiple readings and analyses of those abstracts. Those interested in an overall impression of the research topics in a certain field, as was our purpose in this review, may find this approach useful especially if the number of studies to analyze is substantial. We also searched for subtopics within the automatically-generated topics to get a more fine-grained understanding of what researchers in the field are interested in, but dividing papers into subtopics was easier when knowing they all share an overarching theme.

However, this approach is not without its limitations. A notable one is that some of the studies were not categorized by the topic model as one would expect. We will discuss an illustrative example that demonstrates the limitations of topic modeling for literature reviews. Our model placed Mayerl’s (2005) study on respondents’ baseline speed into the aberrant test-taking behavior group, and not the response time-general group as one would expect given that no aberrant behaviors were at the center of the study (many studies had the opposite issue, as discussed in “[Aberrant test-taking behaviors](#)” Section). Looking at the probabilities associated with this paper, it indeed looks like this paper had rather similar probabilities of being generated from many groups: it had a probability of 0.26 to be generated from aberrant test-taking behavior, 0.22 to be generated from response time-general, but also a probability of 0.24 and 0.20 to be generated from response time models and CPS, respectively. Evidently, this paper’s categorization was not clear-cut. Examining the paper’s abstract suggests that the topic of aberrant test-taking behavior may have been preferred because this abstract included words like “speed” which is associated with speededness, and “bias” which is a possible result of aberrant behavior during testing.

This demonstrates how the use of specific words could drastically impact where documents are placed, a serious limitation if one wishes to assign papers into groups and study trends within these groups (as was reported for example by Chen et al., 2022). Relatively minor linguistic changes or small differences in probabilities could lead to drastic differences in where a paper is assigned. This may lead to a substantial impact

when trying to compare the topics' sizes or changes in these topics' popularity in different years or platforms. So, although we provided information on the topics' sizes for the interested reader, we suggest that topic modeling should be used mostly for identifying themes within a body of literature rather than for comparing how many studies were related to each topic.

Beyond the built-in limitations of topic modeling for literature reviews, our specific application involved a relatively small dataset. While hundreds of documents can be used in a topic model, much larger datasets are usually required for a stable model (Zhao et al., 2015). However, since we used topic modeling as an initial way of extracting repeating topics in the field and were less interested in categorizing our documents, we believe topic modeling is appropriate and useful, as demonstrated by our rather interpretable results.

In addition to the limitation related to topic modeling, this study had other limitations. For example, since we wanted to be as comprehensive as possible, we included any relevant empirical study we could find. This resulted in several issues. First, we did not limit ourselves to peer-reviewed publications nor did we evaluate the quality of the studies. While it is our impression that the studies were of very high quality, we did not formally examine this assertion. Second, we included book chapters and technical reports that sometimes include data and analyses that were also presented in other venues like journal articles. However, since our focus was not on how many studies fell under each topic or subtopic, we feel that this had a minimal effect on our findings.

Conclusion

In this work, we used topic modeling to identify common themes in the literature using process data in LSAs. We identified six groups: response time models, response time-general, aberrant test-taking behavior, action sequences, CPS, and digital writing. Based on these groups, we suggested several important new directions for future research, including research using different types of process data and research in reading and math. We hope that our work will inspire other researchers to pursue these and other directions not suggested here. Doing so can improve our understanding of test-takers' response processes, thus enhancing the validity of LSAs as well as what is known about how people solve problems more broadly.

Appendix A

The identified paper and their topic probabilities.

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Almond et al. (2012)	0.09	0.04	0.01	0.37	0.01	0.49
Araneda et al. (2022)	0.01	0.62	0.09	0.06	0.21	0.01
Ardington et al. (2021)	0.01	0.01	0.01	0.48	0.07	0.43
Aristizábal (2018)	0.01	0.01	0.04	0.79	0.04	0.11
Arslan et al. (2020)	0.01	0.85	0.02	0.02	0.10	0.01
Attali (2016)	0.00	0.02	0.92	0.01	0.05	0.00
Azzolini et al. (2019)	–					

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Barkaoui (2016)	0.00	0.00	0.00	0.00	0.00	0.99
Barkaoui (2019)	0.00	0.00	0.00	0.07	0.00	0.91
Bennett et al. (2007)	–					
Bennett et al. (2010)	0.01	0.04	0.02	0.64	0.28	0.01
Bennett et al. (2020)	0.00	0.01	0.00	0.01	0.00	0.98
Bennett et al. (2021)	0.00	0.00	0.00	0.01	0.00	0.98
Bennett et al. (2022)	0.01	0.02	0.02	0.12	0.01	0.82
Bergstrom et al. (1994)	0.03	0.94	0.02	0.00	0.00	0.00
Bolsinova et al. (2017a, 2017b)	0.72	0.23	0.04	0.00	0.00	0.00
Bolsinova et al. (2017a, 2017b)	0.98	0.01	0.00	0.00	0.00	0.00
Bolsinova and Maris (2016)	0.97	0.01	0.01	0.01	0.00	0.00
Bolsinova and Tijmstra (2016)	0.96	0.01	0.01	0.01	0.00	0.00
Bolsinova and Tijmstra (2018)	0.99	0.00	0.00	0.00	0.00	0.00
Bridgeman and Cline (2000)	0.01	0.47	0.51	0.00	0.01	0.00
Bridgeman and Cline (2004)	0.02	0.49	0.46	0.01	0.02	0.00
Cao et al. (2020)	0.00	0.00	0.00	0.04	0.00	0.95
Chan (2017)	0.01	0.01	0.00	0.15	0.05	0.78
Chang et al. (2005)	0.01	0.64	0.33	0.00	0.01	0.00
Chang et al. (2011)	0.01	0.39	0.59	0.00	0.00	0.00
Chen et al. (2018)	0.02	0.64	0.05	0.01	0.28	0.00
Chen et al. (2019)	0.00	0.01	0.00	0.24	0.75	0.00
Choe et al. (2018)	0.76	0.20	0.03	0.01	0.00	0.00
Chukharev-Hudilainen et al. (2019)	0.10	0.01	0.01	0.03	0.00	0.85
Crawford et al. (2008)	0.00	0.00	0.00	0.01	0.00	0.98
Cui et al. (2020)	0.01	0.00	0.00	0.95	0.03	0.01
Davison et al. (2012)	0.28	0.52	0.03	0.02	0.01	0.13
De Boeck et al. (2017)	0.89	0.10	0.01	0.01	0.00	0.00
Deane (2014)	0.00	0.01	0.00	0.06	0.01	0.92
Deane, O'Reilly, et al. (2018a)	0.01	0.01	0.03	0.15	0.01	0.78
Deane, Roth, et al. (2018b)	0.00	0.00	0.00	0.00	0.00	0.99
Deane and Zhang (2015)	0.04	0.03	0.01	0.02	0.00	0.90
DeCarlo (2021)	0.86	0.10	0.01	0.01	0.00	0.00
Eichmann et al. (2019)	0.00	0.01	0.00	0.02	0.96	0.00
Eichmann et al. (2020a, 2020b)	0.00	0.01	0.02	0.17	0.80	0.00
Eichmann et al. (2020a, 2020b)	0.00	0.02	0.01	0.04	0.92	0.00
Engelhardt and Goldhammer (2019)	0.01	0.02	0.05	0.05	0.87	0.00
Ercikan et al. (2020)	0.08	0.06	0.31	0.52	0.01	0.01
Fox and Marianti (2017)	0.97	0.01	0.01	0.01	0.00	0.00
Fu and Wise (2012)	0.11	0.28	0.15	0.02	0.01	0.43
Glas and van der Linden (2010)	0.98	0.01	0.01	0.00	0.00	0.00
Goldhammer et al. (2013)	0.01	0.15	0.01	0.07	0.74	0.02
Goldhammer et al. (2014)	0.00	0.01	0.01	0.01	0.97	0.00
Goldhammer et al. (2016)	0.01	0.01	0.92	0.01	0.06	0.00
Goldhammer et al. (2017a, 2017b)	0.03	0.28	0.40	0.02	0.26	0.00
Goldhammer, Naumann, et al. (2013)	0.01	0.01	0.03	0.39	0.56	0.01
Goldhammer et al. (2020)	0.00	0.00	0.00	0.98	0.01	0.01

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Goldhammer et al. (2021a, 2021b)	0.13	0.10	0.01	0.01	0.76	0.00
Gong et al. (2020)	0.01	0.38	0.22	0.03	0.36	0.01
Gong et al. (2022)	0.00	0.00	0.00	0.01	0.00	0.98
Gorgun and Bulut (2021)	0.21	0.11	0.65	0.01	0.00	0.01
Greiff et al. (2013)	0.00	0.01	0.01	0.01	0.97	0.00
Greiff et al. (2015)	0.00	0.03	0.01	0.13	0.82	0.00
Greiff et al. (2016)	0.00	0.00	0.00	0.06	0.93	0.00
Guerra-Carrillo and Bunge (2018)	0.02	0.01	0.00	0.02	0.95	0.00
Guo et al. (2018)	0.24	0.01	0.01	0.02	0.00	0.72
Guo et al. (2020)	0.04	0.01	0.00	0.06	0.00	0.88
Hahnel et al. (2016)	0.00	0.00	0.00	0.01	0.98	0.00
Halkitis et al. (1996)	0.04	0.92	0.02	0.00	0.01	0.00
Han et al. (2019)	0.01	0.00	0.00	0.96	0.02	0.01
Hao et al. (2015)	0.01	0.00	0.00	0.80	0.01	0.17
He,Borgonivi et al. (2019)	0.01	0.01	0.01	0.94	0.03	0.01
He, Liao et al. (2019)	0.01	0.01	0.00	0.65	0.33	0.00
He et al. (2018)	0.01	0.01	0.01	0.94	0.03	0.00
He et al. (2021)	0.00	0.01	0.01	0.67	0.30	0.01
He and von Davier (2015)	0.02	0.01	0.01	0.89	0.04	0.04
He and von Davier (2016)	0.01	0.00	0.00	0.92	0.02	0.04
Hecht et al. (2017)	0.05	0.81	0.08	0.02	0.02	0.01
Heldt et al. (2020)	0.01	0.04	0.01	0.11	0.22	0.62
Herborn et al. (2020)	0.02	0.02	0.01	0.16	0.79	0.01
leong et al. (2015)					–	
Ilgun Dibek (2021a)	0.08	0.38	0.01	0.46	0.05	0.01
Ilgun Dibek (2021b)	0.03	0.18	0.47	0.12	0.20	0.01
Ivanova et al. (2020)	0.02	0.33	0.28	0.32	0.04	0.01
Jansen (1997)	0.84	0.12	0.01	0.01	0.01	0.00
Jensen et al. (2018)	0.00	0.01	0.98	0.00	0.00	0.00
Jeon and De Boeck (2019)	0.56	0.09	0.03	0.08	0.24	0.00
Jodoin (2003)	0.02	0.96	0.01	0.01	0.01	0.00
Juškaite (2019)	0.02	0.03	0.01	0.66	0.27	0.02
Kahraman et al. (2013)	0.17	0.09	0.41	0.16	0.16	0.01
Keller et al. (2019)	0.01	0.04	0.01	0.02	0.10	0.83
Khuder and Harwood (2015)	0.04	0.02	0.17	0.01	0.00	0.76
Khuder and Harwood (2019)	0.01	0.10	0.01	0.01	0.00	0.86
Klein Entink et al.(2009a, 2009b)	0.90	0.02	0.01	0.07	0.00	0.00
Kroehne et al. (2019)	0.20	0.06	0.01	0.26	0.46	0.00
Kroehne et al. (2020)	0.01	0.02	0.96	0.01	0.00	0.00
Kroehne and Goldhammer (2018)	0.01	0.01	0.00	0.96	0.01	0.01
Kuhfeld and Soland (2020)	0.01	0.01	0.97	0.00	0.00	0.00
Kupiainen et al. (2014)	0.03	0.02	0.16	0.07	0.71	0.00
Lee and Haberman (2016)	0.03	0.13	0.31	0.45	0.07	0.02
Lee and Jia (2014)	0.08	0.04	0.84	0.03	0.01	0.00
Lee and Wollack (2020)					–	
Lee et al. (2019)	0.02	0.40	0.04	0.39	0.13	0.01

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Li (2021)	0.19	0.01	0.01	0.05	0.00	0.74
Liao et al. (2019)	0.00	0.00	0.00	0.79	0.19	0.01
Lindner et al. (2017)	0.01	0.74	0.19	0.01	0.04	0.01
Lindner et al. (2019)	0.02	0.02	0.65	0.01	0.30	0.00
Ling (2016)	0.02	0.17	0.06	0.02	0.01	0.74
Liu et al. (2018)	0.08	0.01	0.01	0.34	0.57	0.00
Liu et al. (2020)	0.64	0.02	0.33	0.01	0.00	0.00
Los et al. (2020)	0.01	0.03	0.90	0.01	0.05	0.00
Lu et al. (2020)	0.78	0.01	0.20	0.00	0.00	0.00
Maddox et al. (2018)	0.01	0.01	0.00	0.94	0.03	0.01
Man et al. (2019)	0.78	0.04	0.01	0.13	0.03	0.01
Margolis et al. (2020)	0.02	0.80	0.04	0.08	0.01	0.04
Mayerl (2005)	0.24	0.22	0.26	0.05	0.20	0.02
Meng et al. (2015)	0.99	0.00	0.00	0.00	0.00	0.00
Michaelides et al. (2020)	0.03	0.36	0.56	0.01	0.04	0.00
Molenaar and De Boeck (2018)	0.93	0.02	0.01	0.03	0.01	0.00
Mustafić et al. (2019)	0.00	0.01	0.04	0.01	0.92	0.00
Naumann (2015)	0.00	0.00	0.00	0.01	0.98	0.00
Naumann (2019)	0.00	0.00	0.00	0.01	0.99	0.00
Naumann and Goldhammer (2017)	0.00	0.01	0.00	0.01	0.98	0.00
Niepel et al. (2016)	–					
Osman et al. (2017)	0.01	0.20	0.01	0.30	0.04	0.45
Parshall et al. (1994)	0.02	0.96	0.01	0.00	0.00	0.00
Pohl et al. (2019)	0.98	0.01	0.01	0.01	0.00	0.00
Pokropek (2016)	0.66	0.02	0.29	0.02	0.01	0.00
Ponce et al. (2020)	0.00	0.96	0.01	0.00	0.00	0.03
Ponce et al. (2021)	0.00	0.98	0.01	0.00	0.01	0.00
Pools and Monseur (2021)	0.04	0.01	0.93	0.01	0.00	0.00
Qian et al. (2016)	0.02	0.96	0.01	0.00	0.00	0.00
Qian & Jiao (2018)	0.01	0.00	0.00	0.97	0.01	0.01
Rafferty et al. (2020)	0.07	0.06	0.01	0.32	0.51	0.03
Ramalingam and Adams (2018)	0.01	0.01	0.00	0.86	0.12	0.00
Ranger et al. (2020)	0.98	0.01	0.01	0.00	0.00	0.00
Révész et al. (2017)	0.00	0.00	0.00	0.01	0.00	0.98
Révész et al. (2019)	0.00	0.00	0.00	0.00	0.00	0.99
Rios (2021)	0.00	0.01	0.98	0.00	0.00	0.00
Rios and Guo (2020)	0.03	0.20	0.76	0.01	0.01	0.00
Rios and Liu (2017)	0.01	0.02	0.84	0.08	0.01	0.04
Rios and Soland (2021)	0.61	0.01	0.37	0.00	0.00	0.00
Rios et al. (2014)	0.02	0.08	0.87	0.01	0.01	0.00
Rios et al. (2017)	0.01	0.03	0.95	0.00	0.00	0.00
Sahin and Colvin (2020)	0.02	0.38	0.55	0.04	0.01	0.00
Salles et al. (2020)	0.00	0.00	0.00	0.98	0.01	0.00
Saß et al. (2012)	0.01	0.89	0.02	0.01	0.06	0.01
Scherer et al. (2015)	0.00	0.01	0.00	0.01	0.98	0.00
Schnipke and Pashley (1997)	0.15	0.33	0.41	0.06	0.04	0.01
Schnipke and Scrams (1997)	0.11	0.21	0.65	0.01	0.01	0.00
Schnipke and Scrams (1999)	0.45	0.53	0.01	0.01	0.00	0.00

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Setzer et al. (2013)	0.01	0.23	0.75	0.00	0.01	0.00
Shao et al. (2016)	0.86	0.02	0.01	0.09	0.01	0.01
Shin et al. (2020)	0.18	0.61	0.01	0.15	0.04	0.00
Shu et al. (2017)	0.20	0.01	0.00	0.78	0.01	0.00
Sinharay and Johnson (2020)	0.52	0.24	0.07	0.16	0.01	0.01
Sinharay et al. (2019)	0.01	0.01	0.00	0.70	0.02	0.26
Soland (2018a)	0.01	0.00	0.98	0.00	0.00	0.00
Soland (2018b)						
Soland et al., (2019a, 2019b)	0.00	0.01	0.97	0.00	0.01	0.00
Soland and Kuhfeld (2019)	0.00	0.01	0.98	0.00	0.00	0.00
Soland et al., (2019a, 2019b)	0.12	0.04	0.81	0.02	0.01	0.00
Soland et al. (2021)	0.04	0.12	0.73	0.08	0.03	0.01
Stadler et al. (2019)	0.00	0.01	0.00	0.53	0.45	0.01
Stickney et al. (2012)	0.00	0.01	0.01	0.01	0.00	0.97
Su and Davison (2019)	0.03	0.65	0.01	0.01	0.29	0.01
Sukkariéh et al. (2012)	0.02	0.01	0.02	0.88	0.01	0.07
Swanson et al. (2001)	–					
Swanson et al. (2005)	0.01	0.96	0.01	0.00	0.00	0.02
Talebinamvar and Zarrabi (2022)	0.00	0.01	0.02	0.32	0.02	0.63
Talento-Miller (2013)	0.03	0.10	0.84	0.01	0.01	0.01
Tang et al. (2020)	0.01	0.01	0.00	0.96	0.01	0.00
Tang et al. (2021)	0.06	0.01	0.01	0.87	0.04	0.01
Tate and Warschauer (2019)	0.00	0.00	0.00	0.09	0.01	0.89
Tate et al. (2019)	0.00	0.00	0.00	0.01	0.00	0.97
Teig et al. (2020)	0.02	0.01	0.03	0.31	0.61	0.02
Tóth et al. (2017)	0.01	0.01	0.05	0.63	0.30	0.01
Tsaousis et al. (2018)	0.03	0.95	0.01	0.00	0.00	0.00
Ulitzsch et al. (2020a)	0.43	0.01	0.50	0.05	0.00	0.00
Ulitzsch et al. (2020b)	0.95	0.01	0.01	0.02	0.00	0.00
Ulitzsch et al. (2020c)	0.59	0.07	0.02	0.30	0.02	0.01
Ulitzsch et al. (2021)	0.16	0.02	0.01	0.78	0.03	0.00
Ulitzsch et al. (2022)	0.01	0.02	0.14	0.40	0.42	0.01
van der Linden (2006)	0.98	0.01	0.00	0.00	0.00	0.00
van der Linden (2007)	0.98	0.01	0.00	0.00	0.00	0.00
van der Linden (2008)	0.98	0.01	0.01	0.01	0.00	0.00
van der Linden and Glas (2010)	0.95	0.01	0.01	0.02	0.00	0.00
van der Linden and Guo (2008)	0.94	0.02	0.01	0.03	0.00	0.00
van der Linden and van Krimpen-Stoop (2003)	0.98	0.01	0.01	0.00	0.00	0.00
van der Linden et al. (1999)	0.75	0.19	0.04	0.01	0.00	0.00
van der Linden et al. (2007)	0.82	0.14	0.02	0.01	0.00	0.00
van Rijn and Ali (2017)	0.86	0.02	0.05	0.02	0.00	0.04
van Rijn et al. (2021)	0.27	0.42	0.04	0.01	0.01	0.26
von Davier et al. (2019)	0.05	0.02	0.04	0.84	0.02	0.03
Vörös and Rouet (2016)	0.01	0.01	0.01	0.47	0.50	0.01
Wan and Henly (2012)	0.01	0.97	0.01	0.00	0.00	0.00
Wang et al., (2013a, 2013b)	0.99	0.00	0.00	0.00	0.00	0.00
Wang et al., (2013a, 2013b)	0.99	0.00	0.00	0.00	0.00	0.00

Paper	Response time models	Response time-general	Aberrant test-taking behaviors	Action sequences	CPS	Digital writing
Wang and Hanson (2005)	0.99	0.00	0.00	0.00	0.00	0.00
Wang and Xu (2015)	0.57	0.01	0.42	0.00	0.00	0.00
Wang et al., (2018a, 2018b)	0.87	0.01	0.11	0.01	0.00	0.00
Wang et al. (2018a, 2018b)	0.52	0.18	0.28	0.01	0.00	0.00
Weeks et al. (2016)	0.29	0.42	0.26	0.02	0.02	0.00
Wise (2014)	0.01	0.03	0.95	0.00	0.00	0.00
Wise (2015)	0.02	0.02	0.94	0.00	0.00	0.00
Wise (2017)	0.01	0.14	0.84	0.00	0.00	0.00
Wise (2019)	0.15	0.02	0.81	0.01	0.00	0.00
Wise (2020)	0.01	0.37	0.56	0.01	0.04	0.00
Wise and Gao (2017)	0.03	0.32	0.51	0.12	0.01	0.01
Wise and Kingsbury (2016)	0.30	0.04	0.64	0.01	0.01	0.00
Wise and Kuhfeld (2020)	–					
Wise and Kuhfeld (2021a)	0.01	0.01	0.97	0.00	0.00	0.00
Wise and Kuhfeld (2021b)	0.01	0.01	0.98	0.00	0.00	0.00
Wise and Ma (2012)	0.02	0.07	0.90	0.01	0.00	0.00
Wise et al. (2004)	–					
Wise et al. (2005)	–					
Wise et al. (2010)	0.01	0.05	0.93	0.00	0.01	0.00
Wise et al. (2012)	0.05	0.02	0.74	0.17	0.02	0.01
Wise et al. (2019)	0.01	0.01	0.97	0.01	0.00	0.00
Wise et al. (2020)	0.00	0.01	0.98	0.00	0.00	0.00
Wise et al., (2021a, 2021b)	0.00	0.01	0.98	0.00	0.00	0.00
Wise et al., (2021a, 2021b)	0.01	0.54	0.39	0.04	0.02	0.01
Wüstenberg et al. (2014)	0.00	0.00	0.00	0.01	0.99	0.00
Xu et al. (2018)	0.13	0.01	0.01	0.58	0.27	0.00
Yamamoto and Lennon (2018)	0.01	0.01	0.06	0.90	0.02	0.01
Yavuz (2019)	0.02	0.38	0.01	0.28	0.30	0.01
Zehner et al. (2020)	0.03	0.24	0.01	0.26	0.22	0.24
Zenisky and Baldwin (2006)	0.01	0.84	0.02	0.02	0.11	0.01
Zhan et al. (2018)	0.97	0.01	0.00	0.01	0.00	0.00
Zhang et al. (2019)	0.00	0.00	0.00	0.01	0.00	0.98
Zhang and Deane (2015)	0.00	0.00	0.00	0.01	0.00	0.99
Zhang et al. (2016)	0.01	0.01	0.00	0.20	0.01	0.77
Zhang, Zhu, et al. (2017)	0.00	0.01	0.03	0.02	0.00	0.94
Zhang, Zou, et al. (2017)	–					
Zhu et al. (2016)	0.01	0.00	0.00	0.96	0.02	0.01
Zhu et al. (2019)	0.01	0.02	0.14	0.25	0.03	0.55

The topic with highest probability to which the paper was assigned is bolded.—refers to identified papers without an abstract which were not included in the topic model

Acknowledgements

Not applicable.

Author contributions

EA: conceptualization, methodology, formal analysis, investigation, writing—original draft, LK: writing—review and editing, supervision, MvD: resources, writing—review and editing.

Funding

Our work was funded by TIMSS 2019 and PIRLS 2021.

Availability of data and materials

The reviewed paper and the analysis results are provided as an appendix to the paper.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 28 June 2023 Accepted: 25 April 2024

Published online: 06 May 2024

References

- Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report Series*, 2012, i–61.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA, APA, & NCME]. (2014). *Standards for educational and psychological testing*. American Psychological Association.
- Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., & Keehner, M. (2022). Exploring relationships among test takers' behaviors and performance using response process data. *Education Sciences*, 12, 104. <https://doi.org/10.3390/educsci12020104>
- Ardington, C., Wills, G., Pretorius, E., Mohohlwane, N., & Menendez, A. (2021). Benchmarking oral reading fluency in the early grades in Nguni languages. *International Journal of Educational Development*, 84, 102433. <https://doi.org/10.1016/j.ijedudev.2021.102433>
- Aristizábal, J. (2018). Using learning analytics to improve students' reading skills: A case study in an American international school with English as an additional language (EAL) students. *GIST Education and Learning Research Journal*, 17, 193–214.
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19–32. <https://doi.org/10.1080/1364557032000119616>
- Arslan, B., Jiang, Y., Keehner, M., Gong, T., Katz, I. R., & Yan, F. (2020). The effect of drag-and-drop item features on test-taker performance and response strategies. *Educational Measurement Issues and Practice*, 39, 96–106. <https://doi.org/10.1111/emip.12326>
- Attali, Y. (2016). Effort in low-stakes assessments: What does it take to perform as well as in a high-stakes setting? *Educational and Psychological Measurement*, 76, 1045–1058. <https://doi.org/10.1177/0013164416634789>
- Azzolini, D., Bazoli, N., Lievore, I., Schizzerotto, A., & Vergolini, L. (2019). *Beyond achievement. a comparative look into 15 year-olds' school engagement, effort and perseverance in the European Union*. European commission.
- Banihashem, S. K., Aliabadi, K., Pourroostaei Ardakani, S., Delavar, A., & Nili Ahmadabadi, M. (2018). Learning analytics: A systematic literature review. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*. <https://doi.org/10.5812/ijvlms.63024>
- Banihashem, S. K., Aliabadi, K., Pourroostaei Ardakani, S., Nili Ahmadabadi, M. R., & Delavar, A. (2019). Investigation on the role of learning theory in learning analytics. *Interdisciplinary Journal of Virtual Learning in Medical Sciences*, 10(4), 14–27. <https://doi.org/10.30476/IJVLMS.2019.84294.1001>
- Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, 100, 320–340. <https://doi.org/10.1111/modl.12316>
- Barkaoui, K. (2019). What can L2 writers' pausing behavior tell us about their L2 writing process? *Studies in Second Language Acquisition*, 41, 529–554. <https://doi.org/10.1017/S027226311900010X>
- Bennett, R. E., Persky, H., Weiss, A. R., & Jenkins, F. (2007). *Problem solving in technology-rich environments: A report from the NAEP technology based assessment project (NCES 2007–466)*. US Department of Education. National Center for Education Statistics.
- Bennett, R. E., Persky, H., Weiss, A., & Jenkins, F. (2010). Measuring problem solving with technology: A demonstration study for NAEP. *Journal of Technology, Learning, and Assessment*, 8
- Bennett, R. E., Zhang, M., Deane, P., & van Rijn, P. W. (2020). How do proficient and less proficient students differ in their composition processes? *Educational Assessment*, 25, 198–217. <https://doi.org/10.1080/10627197.2020.1804351>
- Bennett, R. E., Zhang, M., & Sinharay, S. (2021). How do educationally at-risk men and women differ in their essay-writing processes? *Chinese/english Journal of Educational Measurement and Evaluation*, 2, 1.
- Bennett, R. E., Zhang, M., Sinharay, S., Guo, H., & Deane, P. (2022). Are there distinctive profiles in examinee essay-writing processes? *Educational Measurement Issues and Practice*, 41, 55–69. <https://doi.org/10.1111/emip.12469>
- Bergner, Y., & von Davier, A. A. (2019). Process data in NAEP: Past, present, and future. *Journal of Educational and Behavioral Statistics*, 44, 706–732. <https://doi.org/10.3102/1076998618784700>
- Bergstrom, B., Gershon, R., & Lunz, M. E. (1994). *Computerized adaptive testing exploring examinee response time using hierarchical linear modeling [Conference presentation]*. The annual meeting of the National Council on Measurement in Education.
- Bischof, J., & Airoldi, E. M. 2012. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on Machine Learning (icml-12)*. 201–208
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>

- Bolsinova, M., de Boeck, P., & Tijmstra, J. (2017a). Modelling conditional dependence between response time and accuracy. *Psychometrika*, *82*, 1126–1148. <https://doi.org/10.1007/s11336-016-9537-6>
- Bolsinova, M., & Maris, G. (2016). A test for conditional independence between response time and accuracy. *British Journal of Mathematical and Statistical Psychology*, *69*, 62–79. <https://doi.org/10.1111/bmsp.12059>
- Bolsinova, M., & Tijmstra, J. (2016). Posterior predictive checks for conditional independence between response time and accuracy. *Journal of Educational and Behavioral Statistics*, *41*, 123–145. <https://doi.org/10.3102/1076998616631746>
- Bolsinova, M., & Tijmstra, J. (2018). Improving precision of ability estimation: Getting more from response times. *British Journal of Mathematical and Statistical Psychology*, *71*, 13–38. <https://doi.org/10.1111/bmsp.12104>
- Bolsinova, M., Tijmstra, J., & Molenaar, D. (2017b). Response moderation models for conditional dependence between response time and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *70*, 257–279. <https://doi.org/10.1111/bmsp.12076>
- Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Bridgeman, B., & Cline, F. (2000). Variations in mean response times for questions on the computer-adaptive GRE® general test: Implications for fair assessment. *ETS Research Report Series*, 2000, i–29.
- Bridgeman, B., & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement*, *41*, 137–148. <https://doi.org/10.1111/j.1745-3984.2004.tb01111.x>
- Brückner, S., & Pellegrino, J. W. (2017). Contributions of response processes analysis to the validation of an assessment of higher education students' competence in business and economics. In B. D. Zumbo & A. M. Huble (Eds.), *Understanding and investigating response processes in validation research* (pp. 31–35). Springer International Publishing.
- Cao, Y., Chen, J., Zhang, M., & Li, C. (2020). Examining the writing processes in scenario-based assessment using regression trees. *ETS Research Report Series*, 2020, 1–16.
- Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia*, *7*, 1–27. <https://doi.org/10.1186/s40468-017-0040-5>
- Chang, S. R., Plake, B. S., & Ferdous, A. A. (2005). *Response times for correct and incorrect item responses on computerized adaptive tests*. The 2005 annual meeting of the American Educational Research Association (AERA).
- Chang, S. R., Plake, B. S., Kramer, G. A., & Lien, S. M. (2011). Development and application of detection indices for measuring guessing behaviors and test-taking effort in computerized adaptive testing. *Educational and Psychological Measurement*, *71*, 437–459. <https://doi.org/10.1177/0013164410385110>
- Chen, H., De Boeck, P., Grady, M., Yang, C. L., & Waldschmidt, D. (2018). Curvilinear dependency of response accuracy on response time in cognitive tests. *Intelligence*, *69*, 16–23. <https://doi.org/10.1016/j.intell.2018.04.001>
- Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). Statistical analysis of complex problem-solving process data: An event history analysis approach. *Frontiers in Psychology*, *10*, 486. <https://doi.org/10.3389/fpsyg.2019.00486>
- Chen, X., Zou, D., & Xie, H. (2022). A decade of learning analytics: Structural topic modeling based bibliometric analysis. *Education and Information Technologies*, *27*, 10517–10561. <https://doi.org/10.1007/s10639-022-11046-z>
- Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*, 135–158. <https://doi.org/10.3102/1076998617723642>
- Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H. H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, *41*, 583–604. <https://doi.org/10.1017/S027226311900007X>
- Crawford, L., Lloyd, S., & Knott, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention*, *33*, 108–119. <https://doi.org/10.1177/1534508407311403>
- Cui, Y., Guo, Q., Leighton, J. P., & Chu, M. W. (2020). Log data analysis with ANFIS: A fuzzy neural network approach. *International Journal of Testing*, *20*, 78–96. <https://doi.org/10.1080/15305058.2018.1551225>
- Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*, 245–263. <https://doi.org/10.1177/0013164411408412>
- De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, *70*, 225–237. <https://doi.org/10.1111/bmsp.12094>
- Deane, P. (2014). Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks. *ETS Research Report Series*, 2014, 1–23.
- Deane, P., O'Reilly, T., Chao, S. F., & Dreier, K. (2018a). Writing processes in short written responses to questions probing prior knowledge. *ETS Research Report Series*, 2018, 1–30.
- Deane, P., Roth, A., Litz, A., Goswami, V., Steck, F., Lewis, M., & Richter, T. (2018b). Behavioral differences between retyping, drafting, and editing: A writing process analysis. *ETS Research Memorandum Series*, *109*, 18–26.
- Deane, P., Sabatini, J. S., Feng, G., Sparks, J., Song, Y., Fowles, M., & Foley, C. (2015). Key practices in the English Language Arts (ELA): Linking learning theory, assessment, and instruction (RR-15–17). *ETS Research Report*. <https://doi.org/10.1002/ets2.12063>
- Deane, P., & Zhang, M. (2015). Exploring the feasibility of using writing process features to assess text production skills. *ETS Research Report Series*, 2015, 1–16.
- DeCarlo, L. T. (2021). On joining a signal detection choice model with response time models. *Journal of Educational Measurement*, *58*, 438–464. <https://doi.org/10.1111/jedm.12300>
- Eichmann, B., Goldhammer, F., Greiff, S., Brandhuber, L., & Naumann, J. (2020a). Using process data to explain group differences in complex problem solving. *Journal of Educational Psychology*, *112*, 1546–1562. <https://doi.org/10.1037/edu0000446>
- Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020b). Exploring behavioural patterns during complex problem-solving. *Journal of Computer Assisted Learning*, *36*, 933–956. <https://doi.org/10.1111/jcal.12451>
- Eichmann, B., Goldhammer, F., Greiff, S., Pucite, L., & Naumann, J. (2019). The role of planning in complex problem solving. *Computers & Education*, *128*, 1–12. <https://doi.org/10.1016/j.compedu.2018.08.004>

- Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology, 10*, 1131. <https://doi.org/10.3389/fpsyg.2019.01131>
- Ericikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment, 25*, 179–197. <https://doi.org/10.1080/10627197.2020.1804353>
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*, 223–241. <https://doi.org/10.1177/1745691612460685>
- Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database* (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2019/international-database/>
- Fox, J. P., & Mariani, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement, 54*, 243–262. <https://doi.org/10.1111/jedm.12143>
- Fu, J., & Wise, M. (2012). Statistical report of 2011 CBAL™ multistate administration of reading and writing tests. *ETS Research Report Series, 2012*, i–59.
- Glas, C. A., & van der Linden, W. J. (2010). Marginal likelihood inference for a model for item responses and response times. *British Journal of Mathematical and Statistical Psychology, 63*, 603–626.
- Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analysing log file data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Large-Scale Cognitive Assessment* (pp. 239–269). Springer.
- Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021a). 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., Kroehne, U., Hahnel, C., & De Boeck, P. (2021b). Controlling speed in component skills of reading improves the explanation of reading comprehension. *Journal of Educational Psychology, 113*(5), 861–878. <https://doi.org/10.1037/edu0000655>
- Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC*. OECD Education Working Papers.
- Goldhammer, F., Martens, T., & Lüdtke, O. (2017a). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education, 5*, 1–25. <https://doi.org/10.1186/s40536-017-0051-9>
- Goldhammer, F., Naumann, J., & Keßel, Y. (2013). Assessing individual differences in basic computer skills: Psychometric characteristics of an interactive performance measure. *European Journal of Psychological Assessment, 29*, 263–275. <https://doi.org/10.1027/1015-5759/a000153>
- Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017b). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence Assessment in Education*. Springer.
- Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology, 106*, 608–626. <https://doi.org/10.25656/01:17967>
- Goldhammer, F., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives, 15*, 128–132. <https://doi.org/10.1080/15366367.2017.1411651>
- Gong, T., Shuai, L., Arslan, B., & Jiang, Y. J. (2020). analysis on scientific inquiry tasks using large-scale national assessment dataset. In A. N. Rafferty, J. Whitehill, V. Cavalli-Sforza, & C. Romero (Eds.), *Proceedings of the 13th international conference on educational data mining EDM* (pp. 417–423). Eric.
- Gong, T., Zhang, M., & Li, C. (2022). Association of keyboarding fluency and writing performance in online-delivered assessment. *Assessing Writing, 51*, 100575. <https://doi.org/10.1016/j.asw.2021.100575>
- Gorgun, G., & Bulut, O. (2021). A polytomous scoring approach to handle not-reached items in low-stakes assessments. *Educational and Psychological Measurement, 81*, 847–871. <https://doi.org/10.1177/0013164421991211>
- 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>
- Greiff, S., Wüstenberg, S., Molnár, G., Fischer, A., Funke, J., & Csapó, B. (2013). Complex problem solving in educational contexts-something beyond g: Concept, assessment, measurement invariance, and construct validity. *Journal of Educational Psychology, 105*, 364–379. <https://doi.org/10.1037/a0031856>
- Guerra-Carrillo, B. C., & Bunge, S. A. (2018). Eye gaze patterns reveal how reasoning skills improve with experience. *Npj Science of Learning, 3*, 1–9.
- Guo, H., Deane, P. D., van Rijn, P. W., Zhang, M., & Bennett, R. E. (2018). Modeling basic writing processes from keystroke logs. *Journal of Educational Measurement, 55*, 194–216. <https://doi.org/10.1111/jedm.12172>
- Guo, H., Zhang, M., Deane, P., & Bennett, R. (2020). Effects of scenario-based assessment on students' writing processes. *Journal of Educational Data Mining, 12*, 19–45. <https://doi.org/10.5281/zenodo.3911797>
- Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior, 55*, 486–500. <https://doi.org/10.1016/j.chb.2015.09.042>
- Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996). *Estimating testing time: The effects of item characteristics on response latency*. The Annual Meeting of the American Educational Research Association.
- Han, Z., He, Q., & Von Davier, M. (2019). Predictive feature generation and selection using process data from PISA interactive problem-solving items: An application of random forests. *Frontiers in Psychology, 10*, 2461. <https://doi.org/10.3389/fpsyg.2019.02461>
- Hao, J., Shu, Z., & von Davier, A. (2015). Analyzing process data from game scenario-based tasks: An edit distance approach. *Journal of Educational Data Mining, 7*, 33–50.

- Hayes, J. R. (2012). Modeling and Remodeling Writing. *Written Communication*, 29(3), 369–388. <https://doi.org/10.1177/0741088312451260>
- He, Q., Borgonovi, F., & Paccagnella, M. (2019a). *Using process data to understand adults' problem-solving behaviour in the programme for the international assessment of adult competencies (PIAAC) Identifying generalised patterns across multiple tasks with sequence mining*. OECD Education Working Papers.
- 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.
- He, Q., Liao, D., & Jiao, H. (2019b). Clustering behavioral patterns using process data in PIAAC problem-solving items. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 189–221). Springer.
- He, Q., & von Davier, M. (2015). Identifying feature sequences from process data in problem-solving items with n-grams. In L. A. van der Ark, D. M. Bolt, W. Wang, J. A. Douglas, & S. Chow (Eds.), *Quantitative psychology research* (pp. 173–190). Springer.
- He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams: Insights from a computer-based large-scale assessment. In Y. Rosen, S. Ferrara, & M. Mosharrar (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). IGI Global.
- He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in problem-solving items in computer-based large-scale assessments. In H. Jiao, R. W. Lissitz, & A. Van Wie (Eds.), *Data analytics and psychometrics: informing assessment practices* (pp. 53–76). Information Age Publishing.
- Hecht, M., Siegle, T., & Weirich, S. (2017). A model for the estimation of testlet response time to optimize test assembly in paper-and-pencil large-scale assessments. *Journal for Educational Research Online*, 9, 32–51.
- Heldt, M., Massek, C., Drossel, K., & Eickelmann, B. (2020). The relationship between differences in students' computer and information literacy and response times: An analysis of IEA-ICILS data. *Large-Scale Assessments in Education*, 8, 1–20. <https://doi.org/10.1186/s40536-020-00090-1>
- Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior*, 104, 105624. <https://doi.org/10.1016/j.chb.2018.07.035>
- Jeong, M. K., Sit, P. S., Mak, S. K., & Cheung, K. C. (2015). *Analysis of log file data to understand problem-solving behaviour: An example of a released item from PISA 2012 study*. The 2015 Global Chinese Conference on Educational Information and Assessment & Chinese Association of Psychological Testing 2015 Annual Conference.
- Ilgun Dibek, M. (2021a). A high-stakes approach to response time effort in low-stakes assessment. *International Journal of Educational Methodology*, 7, 571–586. <https://doi.org/10.12973/ijem.7.4.571>
- Ilgun Dibek, M. (2021b). Silent predictors of test disengagement in PIAAC 2012. *Journal of Measurement and Evaluation in Education and Psychology*, 11, 430–450. <https://doi.org/10.21031/epod.796626>
- Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation*, 26, 252–274. <https://doi.org/10.1080/13803611.2021.1963939>
- Jansen, M. G. (1997). The Rasch model for speed tests and some extensions with applications to incomplete designs. *Journal of Educational and Behavioral Statistics*, 22, 125–140. <https://doi.org/10.3102/10769986022002125>
- Jeon, M., & De Boeck, P. (2019). An analysis of an item-response strategy based on knowledge retrieval. *Behavior Research Methods*, 51, 697–719. <https://doi.org/10.3758/s13428-018-1064-1>
- Jensen, N., Rice, A., & Soland, J. (2018). The influence of rapidly guessed item responses on teacher value-added estimates: Implications for policy and practice. *Educational Evaluation and Policy Analysis*, 40, 267–284. <https://doi.org/10.3102/0162373718759600>
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40, 1–15. <https://doi.org/10.1111/j.1745-3984.2003.tb01093.x>
- Juškaite, L. (2019). Data mining in education: Online testing in Latvian schools. In V. Lamanaukas (Ed), *Science and technology education: Current challenges and possible solutions, Proceedings of the 3rd International Baltic Symposium on Science and Technology in Education*. Šiauliai. <https://www.ceeol.com/search/chapter-detail?id=942447>
- Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, 37, 343–360. <https://doi.org/10.1177/0146621613477236>
- Kane, M., & Mislevy, R. (2017). Validating score interpretations based on response processes. *Validation of score meaning for the next generation of assessments* (pp. 11–24). Routledge.
- Keller, J., Ruthruff, E., & Keller, P. (2019). Mindfulness and speed testing for children with learning disabilities: Oil and water? *Reading Writing Quarterly*, 35, 154–178. <https://doi.org/10.1080/10573569.2018.1524803>
- Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: theories, methods, individual differences, and applications* (pp. 57–71). Lawrence Erlbaum Associates Inc.
- Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, 114(2), 175–191. <https://doi.org/10.2307/1423513>
- Khalil, M., Prinsloo, P., & Slade, S. (2023). The use and application of learning theory in learning analytics: A scoping review. *Journal of Computing in Higher Education*. <https://doi.org/10.1007/s12528-022-09340-3>
- Khorramdel, L., von Davier, M., Kirsch, I., & Yamamoto, K. (2023). Educational surveys: conceptual overview. In R. J. Tierney, F. Rizvi, & K. Ericikan (Eds.), *International encyclopedia of education (Fourth Edition)* (pp. 347–358). Elsevier.
- Khuder, B., & Harwood, N. (2015). L2 writing in test and non-test situations: Process and product. *Journal of Writing Research*, 6, 233–278.
- Khuder, B., & Harwood, N. (2019). L2 writing task representation in test-like and non-test-like situations. *Written Communication*, 36, 578–632. <https://doi.org/10.1177/0741088319862779>
- Klein Entink, R. H., Fox, J. P., & van der Linden, W. J. (2009a). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, 74, 21–48. <https://doi.org/10.1007/S11336-008-9075-Y>
- Klein Entink, R., van der Linden, W. J., & Fox, J. P. (2009b). A Box-Cox normal model for response times. *British Journal of Mathematical and Statistical Psychology*, 62, 621–640. <https://doi.org/10.1348/000711008X374126>

- Knight, S., & Shum, S. B. (2017). Theory and learning analytics. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 17–22). Solar.
- Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, 62, 147–177.
- 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, 527–563. <https://doi.org/10.1007/s41237-018-0063-y>
- Kroehne, U., Hahnel, C., & Goldhammer, F. (2019). Invariance of the response processes between gender and modes in an assessment of reading. *Frontiers in Applied Mathematics and Statistics*, 5, 2. <https://doi.org/10.3389/fams.2019.00002>
- Kuhfeld, M., & Soland, J. (2020). Using assessment metadata to quantify the impact of test disengagement on estimates of educational effectiveness. *Journal of Research on Educational Effectiveness*, 13, 147–175. <https://doi.org/10.1080/19345747.2019.1636437>
- Kupiainen, S., Vainikainen, M. P., Marjanen, J., & Hautamäki, J. (2014). The role of time on task in computer-based low-stakes assessment of cross-curricular skills. *Journal of Educational Psychology*, 106, 627–638. <https://doi.org/10.1037/a0035507>
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323. [https://doi.org/10.1016/0010-0285\(74\)90015-2](https://doi.org/10.1016/0010-0285(74)90015-2)
- Lee, Y. H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, 16, 240–267. <https://doi.org/10.1080/15305058.2015.1085385>
- Lee, Y. H., Hao, J., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? *A Response-Time Perspective Frontiers in Psychology*, 10, 906. <https://doi.org/10.3389/fpsyg.2019.00906>
- Lee, Y. H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, 2, 1–24. <https://doi.org/10.1186/s40536-014-0008-1>
- Lee, S., & Wollack, J. A. (2020). Concurrent use of response time and response accuracy for detecting examinees with item preknowledge. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 165–175). Routledge.
- Li, T. (2021). Identifying mixture components from large-scale keystroke log data. *Frontiers in Psychology*, 12, 628660. <https://doi.org/10.3389/fpsyg.2021.628660>
- Liao, D., He, Q., & Jiao, H. (2019). Mapping background variables with sequential patterns in problem-solving environments: An investigation of United States adults' employment status in PIAAC. *Frontiers in Psychology*, 10, 646. <https://doi.org/10.3389/fpsyg.2019.00646>
- Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology*, 51, 482–492. <https://doi.org/10.1016/j.cedpsych.2017.09.009>
- Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology*, 10, 1533. <https://doi.org/10.3389/fpsyg.2019.01533>
- Ling, G. (2016). Does it matter whether one takes a test on an iPad or a desktop computer? *International Journal of Testing*, 16, 352–377. <https://doi.org/10.1080/15305058.2016.1160097>
- Liu, Y., Cheng, Y., & Liu, H. (2020). Identifying effortful individuals with mixture modeling response accuracy and response time simultaneously to improve item parameter estimation. *Educational and Psychological Measurement*, 80, 775–807. <https://doi.org/10.1177/0013164419895068>
- Liu, H., Liu, Y., & Li, M. (2018). Analysis of process data of PISA 2012 computer-based problem solving: Application of the modified multilevel mixture IRT model. *Frontiers in Psychology*, 9, 1372. <https://doi.org/10.3389/fpsyg.2018.01372>
- Los, J. E., Witmer, S. E., & Roseth, C. J. (2020). Exploring motivational factors associated with test-taking effort among middle school students. *School Psychology Review*. <https://doi.org/10.1080/2372966X.2020.1832862>
- Lu, J., Wang, C., Zhang, J., & Tao, J. (2020). A mixture model for responses and response times with a higher-order ability structure to detect rapid guessing behaviour. *British Journal of Mathematical and Statistical Psychology*, 73, 261–288. <https://doi.org/10.1111/bmsp.12175>
- Maddox, B., Bayliss, A. P., Fleming, P., Engelhardt, P. E., Edwards, S. G., & Borgonovi, F. (2018). Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment. *European Journal of Psychology of Education*, 33, 543–558. <https://doi.org/10.1007/s10212-018-0380-2>
- Man, K., Harring, J. R., Jiao, H., & Zhan, P. (2019). Joint modeling of compensatory multidimensional item responses and response times. *Applied Psychological Measurement*, 43, 639–654. <https://doi.org/10.1177/0146621618824853>
- Margolis, M. J., von Davier, M., & Clauser, B. E. (2020). Timing considerations for performance assessments. *Integrating timing considerations to improve testing practices* (pp. 90–103). Routledge.
- Mayerl, J. (2005). Controlling the baseline speed of respondents: An empirical evaluation of data treatment methods of response latencies. *Proceedings of the sixth international conference on logic and methodology* (pp. 1–20). Princeton.
- Meng, X. B., Tao, J., & Chang, H. H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, 52, 1–27. <https://doi.org/10.1111/jedm.12060>
- Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing*, 20, 187–205. <https://doi.org/10.1080/15305058.2019.1706529>
- Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 262–272). Association for Computational Linguistics.
- Molenaar, D., & de Boeck, P. (2018). Response mixture modeling: Accounting for heterogeneity in item characteristics across response times. *Psychometrika*, 83, 279–297. <https://doi.org/10.1007/s11336-017-9602-9>
- Mustafić, M., Yu, J., Stadler, M., Vainikainen, M. P., Bornstein, M. H., Putnick, D. L., & Greiff, S. (2019). Complex problem solving: Profiles and developmental paths revealed via latent transition analysis. *Developmental Psychology*, 55, 2090–2101. <https://doi.org/10.1037/dev0000764>

- Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, 53, 263–277. <https://doi.org/10.1016/j.chb.2015.06.051>
- Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology*, 10, 1429. <https://doi.org/10.3389/fpsyg.2019.01429>
- Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53, 1–16. <https://doi.org/10.1016/j.lindif.2016.10.002>
- Niepel, C., Greiff, S., Scherer, R., & Martin, R. (2016). Using behavioral data from computer-generated log files to understand complex problem solving performance in a computer-based assessment. In *31st International Congress of Psychology (ICP)*. <https://orbilu.uni.lu/handle/10993/24241>
- Oranje, A., Gorin, J., Jia, Y., Kerr, D., Ercikan, K., & Pellegrino, J. W. (2017). Collecting, analysing, and interpreting response time, eye tracking and log data. In K. Erickson & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 39–51). National Council on Measurement in Education.
- Organisation for Economic Co-Operation and Development. (2012). Assessing problem-solving skills in PISA. In M. Achiron (Ed.), *PISA 2012 RESULTS: creative problem solving* (Vol. V, pp. 25–46). London: OECD.
- Osman, M. J., Idris, N. H., Idris, N. H., & Ishak, M. H. I. (2017). Paper versus screen: Assessment of basic literacy skill of Indigenous people. *Advanced Science Letters*, 23, 8752–8757. <https://doi.org/10.1166/asl.2017.9964>
- Pandur, M. B., Dobša, J., & Kronegger, L. (2020). Topic modelling in social sciences: Case study of web of science. *Central European Conference on Information and Intelligent Systems*.
- Parshall, C. G., Mittelholtz, D. J., & Miller, T. R. (1994). *Response latency: An investigation into determinants of item-level timing*. The Annual Meeting of the National Council on Measurement in Education.
- Pohl, S., Ulitzsch, E., & von Davier, M. (2019). Using response times to model not-reached items due to time limits. *Psychometrika*, 84, 892–920. <https://doi.org/10.1007/s11336-019-09669-2>
- Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science*. <https://doi.org/10.1126/science.abd3300>
- Pokropek, A. (2016). Grade of membership response time model for detecting guessing behaviors. *Journal of Educational and Behavioral Statistics*, 41, 300–325. <https://doi.org/10.3102/1076998616636618>
- Ponce, H. R., Mayer, R. E., & Loyola, M. S. (2021). Effects on test performance and efficiency of technology-enhanced items: An analysis of drag-and-drop response interactions. *Journal of Educational Computing Research*, 59, 713–739. <https://doi.org/10.1177/0735633120969666>
- Ponce, H. R., Mayer, R. E., Sitthiworachart, J., & López, M. J. (2020). Effects on response time and accuracy of technology-enhanced cloze tests: An eye-tracking study. *Educational Technology Research and Development*, 68, 2033–2053. <https://doi.org/10.1007/s11423-020-09740-1>
- Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-Scale Assessments in Education*, 9, 1–31. <https://doi.org/10.1186/s40536-021-00104-6>
- Provasnik, S. (2021). Process data, the new frontier for assessment development: Rich new soil or a quixotic quest? *Large-Scale Assessments in Education*, 9(1), 1–17. <https://doi.org/10.1186/s40536-020-00092-z>
- Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. *Educational Measurement Issues and Practice*, 35, 38–47. <https://doi.org/10.1111/emip.12102>
- Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: A didactic. *Frontiers in Psychology*, 9, 2231. <https://doi.org/10.3389/fpsyg.2018.02231>
- Rafferty, A. N., Jansen, R. A., & Griffiths, T. L. (2020). Assessing mathematics misunderstandings via bayesian inverse planning. *Cognitive Science*, 44, e12900. <https://doi.org/10.1111/cogs.12900>
- Ramalingam, D., & Adams, R. J. (2018). How can the use of data from computer-delivered assessments improve the measurement of twenty-first century skills? In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills* (pp. 225–238). Cham: Springer.
- Ranger, J., Kuhn, J. T., & Ortner, T. M. (2020). Modeling responses and response times in tests with the hierarchical model and the three-parameter lognormal distribution. *Educational and Psychological Measurement*, 80, 1059–1089. <https://doi.org/10.1177/0013164420908916>
- Reis Costa, D., & Leoncio Netto, W. (2022). Process data analysis in ILSAs. In T. Nilsen, A. Stancel-Piątak, & J. E. Gustafsson (Eds.), *International Handbook of Comparative Large-Scale Studies in Education* (pp. 1–27). Cham: Springer.
- Révész, A., Michel, M., & Lee, M. (2017). Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory. *IELTS Research Reports Online Series*.
- Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition*, 41, 605–631. <https://doi.org/10.1017/S027226311900024X>
- Rios, J. A. (2021). Is differential noneffortful responding associated with type I error in measurement invariance testing? *Educational and Psychological Measurement*, 81, 957–979. <https://doi.org/10.1177/0013164421990429>
- Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33, 263–279. <https://doi.org/10.1080/08957347.2020.1789141>
- Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing*, 17, 74–104. <https://doi.org/10.1080/15305058.2016.1231193>
- Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, 31, 226–241. <https://doi.org/10.1080/08923647.2017.1258628>
- Rios, J. A., Liu, O. L., & Bridgeman, B. (2014). Identifying low-effort examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*, 2014, 69–82. <https://doi.org/10.1002/ir.20068>

- Rios, J. A., & Soland, J. (2021). Parameter estimation accuracy of the Effort-Moderated Item Response Theory Model under multiple assumption violations. *Educational and Psychological Measurement, 81*, 569–594. <https://doi.org/10.1177/0013164420949896>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm An R package for structural topic models. *Journal of Statistical Software, 91*, 1–40. <https://doi.org/10.18637/jss.v000.i00>
- Rupp, A. A., Gushta, M., Mislevy, R. J., & Shaffer, D. W. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *The Journal of Technology Learning and Assessment, 8*(4), e1623.
- Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988–1003. <https://doi.org/10.1080/01621459.2016.1141684>
- Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education, 8*, 1–24. <https://doi.org/10.1186/s40536-020-00082-1>
- Salles, F., Dos Santos, R., & Keskaik, S. (2020). When didactics meet data science: Process data analysis in large-scale mathematics assessment in France. *Large-Scale Assessments in Education, 8*, 1–20. <https://doi.org/10.1186/s40536-020-00085-y>
- Saß, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: Effects on response time and response correctness. *Applied Cognitive Psychology, 26*, 70–81. <https://doi.org/10.1002/acp.1798>
- Scherer, R., Greiff, S., & Hautamäki, J. (2015). Exploring the relation between time on task and ability in complex problem solving. *Intelligence, 48*, 37–50. <https://doi.org/10.1016/j.intell.2014.10.003>
- Schnipke, D., & Pashley, P. (1997). *Assessing subgroup differences in response times*. The Annual Meeting of the American Educational Research Association.
- Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213–232. <https://doi.org/10.1111/j.1745-3984.1997.tb00516.x>
- Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks*. Law School Admission Council Computerized Testing Report. Law School Admission Council.
- Schoenfeld, A. H. (2013). Reflections on problem solving theory and practice. *The Mathematics Enthusiast, 10*, 9–34.
- Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. <https://doi.org/10.1080/08957347.2013.739453>
- Shao, C., Li, J., & Cheng, Y. (2016). Detection of test speededness using change-point analysis. *Psychometrika, 81*, 1118–1141. <https://doi.org/10.1007/s11336-015-9476-7>
- Shin, H. J., Kerzabi, E., Joo, S. H., Robin, F., & Yamamoto, K. (2020). Comparability of response time scales in PISA. *Psychological Test and Assessment Modeling, 62*, 107–135.
- Shu, Z., Bergner, Y., Zhu, M., Hao, J., & von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychological Test and Assessment Modeling, 59*, 109.
- Sinharay, S., & Johnson, M. S. (2020). The use of item scores and response times to detect examinees who may have benefited from item preknowledge. *British Journal of Mathematical and Statistical Psychology, 73*, 397–419. <https://doi.org/10.1111/bmsp.12187>
- Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of essay scores from writing process and product features using data mining methods. *Applied Measurement in Education, 32*, 116–137. <https://doi.org/10.1080/08957347.2019.1577245>
- Soland, J. (2018a). Are achievement gap estimates biased by differential student test effort? putting an important policy metric to the test. *Teachers College Record, 120*, 1–26.
- Soland, J. (2018b). The achievement gap or the engagement gap? Investigating the sensitivity of gaps estimates to test motivation. *Applied Measurement in Education, 31*, 312–323. <https://doi.org/10.1080/08957347.2018.1495213>
- Soland, J., Jensen, N., Keys, T. D., Bi, S. Z., & Wolk, E. (2019a). Are test and academic disengagement related? implications for measurement and practice. *Educational Assessment, 24*, 119–134. <https://doi.org/10.1080/10627197.2019.1575723>
- Soland, J., & Kuhfeld, M. (2019). Do students rapidly guess repeatedly over time? a longitudinal analysis of student test disengagement, background, and attitudes. *Educational Assessment, 24*, 327–342. <https://doi.org/10.1080/10627197.2019.1645592>
- Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education, 9*, 1–21. <https://doi.org/10.1186/s40536-021-00100-w>
- Soland, J., Wise, S. L., & Gao, L. (2019b). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education, 32*, 151–165. <https://doi.org/10.1080/08957347.2019.1577244>
- 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>
- Stickney, E. M., Sharp, L. B., & Kenyon, A. S. (2012). Technology-enhanced assessment of math fact automaticity: Patterns of performance for low-and typically achieving students. *Assessment for Effective Intervention, 37*, 84–94. <https://doi.org/10.1177/1534508411430321>
- Su, S., & Davison, M. L. (2019). Improving the predictive validity of reading comprehension using response times of correct item responses. *Applied Measurement in Education, 32*, 166–182. <https://doi.org/10.1080/08957347.2019.1577247>
- Sukkarieh, J. Z., von Davier, M., & Yamamoto, K. (2012). From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks. *ETS Research Report Series, 2012*, i–43.
- Swanson, D. B., Case, S. M., Ripkey, D. R., Clauser, B. E., & Holtman, M. C. (2001). Relationships among item characteristics, examinee characteristics, and response times on USMLE Step 1. *Academic Medicine, 76*, S114–S116.
- Swanson, D. B., Holtzman, K. Z., Clauser, B. E., & Sawhill, A. J. (2005). Psychometric characteristics and response times for one-best-answer questions in relation to number and source of options. *Academic Medicine, 80*, S93–S96.
- Talebinavar, M., & Zarrabi, F. (2022). Clustering students' writing behaviors using keystroke logging: A learning analytic approach in EFL writing. *Language Testing in Asia, 12*, 1–20. <https://doi.org/10.1186/s40468-021-00150-5>

- Talento-Miller, E., Guo, F., & Han, K. T. (2013). Examining test speededness by native language. *International Journal of Testing*, 13, 89–104. <https://doi.org/10.1080/15305058.2011.653021>
- Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, 85, 378–397. <https://doi.org/10.1007/s11336-020-09708-3>
- Tang, X., Wang, Z., Liu, J., & Ying, Z. (2021). An exploratory analysis of the latent structure of process data via action sequence autoencoders. *British Journal of Mathematical and Statistical Psychology*, 74, 1–33. <https://doi.org/10.1111/bmsp.12203>
- Tate, T. P., & Warschauer, M. (2019). Keypresses and mouse clicks: Analysis of the First national computer-based writing assessment. *Technology, Knowledge and Learning*, 24, 523–543. <https://doi.org/10.1007/s10758-019-09412-x>
- Tate, T. P., Warschauer, M., & Kim, Y. S. G. (2019). Learning to compose digitally: The effect of prior computer use and keyboard activity on NAEP writing. *Reading and Writing*, 32, 2059–2082. <https://doi.org/10.1007/s11145-019-09940-z>
- Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching*, 57, 1400–1429. <https://doi.org/10.1002/tea.21657>
- Tóth, K., Rölke, H., Goldhammer, F., & Barkow, I. (2017). Educational process mining: New possibilities for understanding students' problem-solving skills. In B. Csapó & J. Funke (Eds.), *The nature of problem solving: using research to inspire 21st century learning* (pp. 193–209). OECD.
- Tsaousis, I., Sideridis, G. D., & Al-Sadaawi, A. (2018). An IRT–multiple indicators multiple causes (MIMIC) approach as a method of examining item response latency. *Frontiers in Psychology*, 9, 2177. <https://doi.org/10.3389/fpsyg.2018.02177>
- Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child Development*, 51, 1–10. <https://doi.org/10.2307/1129583>
- Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, 47, 3–35. <https://doi.org/10.3102/10769986211010467>
- Ulitzsch, E., He, Q., Ulitzsch, V., Molter, H., Nichterlein, A., Niedermeier, R., & Pohl, S. (2021). Combining clickstream analyses and graph-modeled data clustering for identifying common response processes. *Psychometrika*, 86, 190–214. <https://doi.org/10.1007/s11336-020-09743-0>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020a). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology*, 73, 83–112. <https://doi.org/10.1111/bmsp.12188>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020b). A multiprocess item response model for not-reached items due to time limits and quitting. *Educational and Psychological Measurement*, 80, 522–547. <https://doi.org/10.1177/0013164419878241>
- Ulitzsch, E., von Davier, M., & Pohl, S. (2020c). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research*, 55, 425–453. <https://doi.org/10.1080/00273171.2019.1643699>
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31, 181–204.
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72, 287–308. <https://doi.org/10.1007/s11336-006-1478-z>
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33, 5–20. <https://doi.org/10.3102/1076998607302626>
- van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, 44, 117–130. <https://doi.org/10.1111/j.1745-3984.2007.00030.x>
- van der Linden, W. J., & Glas, C. A. (2010). Statistical tests of conditional independence between responses and or response times on test items. *Psychometrika*, 75, 120–139. <https://doi.org/10.1007/S11336-009-9129-9>
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73, 365–384. <https://doi.org/10.1007/S11336-007-9046-8>
- van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in computerized adaptive testing. *Applied Psychological Measurement*, 23, 195–210.
- van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, 68, 251–265. <https://doi.org/10.1007/BF02294800>
- van Rijn, P. W., & Ali, U. S. (2017). A comparison of item response models for accuracy and speed of item responses with applications to adaptive testing. *British Journal of Mathematical and Statistical Psychology*, 70, 317–345. <https://doi.org/10.1111/bmsp.12101>
- van Rijn, P. W., Attali, Y., & Ali, U. S. (2021). Impact of scoring instructions, timing, and feedback on measurement: An experimental study. *The Journal of Experimental Education*. <https://doi.org/10.1080/00220973.2021.1969532>
- Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, 94, 101582. <https://doi.org/10.1016/j.is.2020.101582>
- Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89, 98–110. <https://doi.org/10.1016/j.chb.2018.07.027>
- 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, 671–705. <https://doi.org/10.3102/1076998619881789>
- von Davier, M., Tyack, L., & Khorramdel, L. (2023). Scoring graphical responses in TIMSS 2019 using artificial neural networks. *Educational and Psychological Measurement*, 83(3), 556–585. <https://doi.org/10.1177/00131644221098021>
- Vörös, Z., & Rouet, J. F. (2016). Laypersons' digital problem solving: Relationships between strategy and performance in a large-scale international survey. *Computers in Human Behavior*, 64, 108–116. <https://doi.org/10.1016/j.chb.2016.06.018>
- Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25, 58–78. <https://doi.org/10.1080/08957347.2012.635507>

- Wang, C., Chang, H. H., & Douglas, J. A. (2013a). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66, 144–168. <https://doi.org/10.1111/j.2044-8317.2012.02045.x>
- Wang, C., Fan, Z., Chang, H. H., & Douglas, J. A. (2013b). A semiparametric model for jointly analyzing response times and accuracy in computerized testing. *Journal of Educational and Behavioral Statistics*, 38, 381–417. <https://doi.org/10.3102/1076998612461831>
- Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29, 323–339. <https://doi.org/10.1177/0146621605275984>
- Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68, 456–477. <https://doi.org/10.1111/bmsp.12054>
- Wang, C., Xu, G., & Shang, Z. (2018a). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika*, 83, 223–254. <https://doi.org/10.1007/s11336-016-9525-x>
- Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018b). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, 43, 469–501. <https://doi.org/10.3102/1076998618767123>
- Weeks, J. P., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, 58, 671–701.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81. <https://doi.org/10.1006/ceps.1999.1015>
- Wise, S. L. (2014). The utility of adaptive testing in addressing the problem of unmotivated examinees. *Journal of Computerized Adaptive Testing*, 2, 1–17. <https://doi.org/10.7333/1401-02010001>
- Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, 28, 237–252. <https://doi.org/10.1080/08957347.2015.1042155>
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36, 52–61. <https://doi.org/10.1111/emip.12165>
- Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, 32, 325–336. <https://doi.org/10.1080/08957347.2019.1660350>
- Wise, S. L. (2020). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education*, 33, 83–94. <https://doi.org/10.1080/08957347.2020.1732386>
- Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30, 343–354. <https://doi.org/10.1080/08957347.2017.1353992>
- Wise, S. L., Im, S., & Lee, J. (2021a). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment*, 26, 163–174. <https://doi.org/10.1080/10627197.2021.1956897>
- Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, 53, 86–105. <https://doi.org/10.1111/jedm.12102>
- Wise, S. L., & Kuhfeld, M. R. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & A. Feinberg (Eds.), *Integrating timing considerations to improve testing practices* (pp. 150–164). Routledge.
- Wise, S., & Kuhfeld, M. (2021a). A method for identifying partial test-taking engagement. *Applied Measurement in Education*, 34, 150–161. <https://doi.org/10.1080/08957347.2021.1890745>
- Wise, S. L., & Kuhfeld, M. R. (2021b). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, 58, 130–149. <https://doi.org/10.1111/jedm.12275>
- Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education*, 32, 183–192. <https://doi.org/10.1080/08957347.2019.1577248>
- Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004). An investigation of motivation filtering in a statewide achievement testing program. *The Annual Meeting of the National Council on Measurement in Education*.
- Wise, S. L., & Ma, L., (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. (pp. 163–183). *The Annual Meeting of the National Council on Measurement in Education*.
- Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). An investigation of the relationship between time of testing and test-taking effort. *The Annual Meeting of the National Council on Measurement in Education*
- Wise, S. L., Ma, L., & Theaker, R. A. (2012). Identifying non-effortful student behavior on adaptive tests: Implications for test fraud detection. In N. Kingston & A. Clark (Eds.), *Test fraud* (pp. 191–201). Routledge.
- Wise, S. L., Owens, K. M., Yang, S. T., Weiss, B., Kissel, H. L., Kong, X., & Horst, S. J. (2005). An investigation of the effects of self-adapted testing on examinee effort and performance in a low-stakes achievement test. *The Annual Meeting of the National Council on Measurement in Education*.
- Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139–153). APA.
- Wise, S. L., Soland, J., & Bo, Y. (2020). The (non) impact of differential test taker engagement on aggregated scores. *International Journal of Testing*, 20, 57–77. <https://doi.org/10.1080/15305058.2019.1605999>
- Wise, S. L., Soland, J., & Dupray, L. M. (2021b). The impact of technology-enhanced items on test-taker disengagement. *Journal of Applied Testing Technology*, 22, 28–36.
- Wüstenberg, S., Stadler, M., Hautamäki, J., & Greiff, S. (2014). The role of strategy knowledge for the application of strategies in complex problem solving tasks. *Technology, Knowledge and Learning*, 19, 127–146. <https://doi.org/10.1007/s10758-014-9222-8>
- Xu, H., Fang, G., Chen, Y., Liu, J., & Ying, Z. (2018). Latent class analysis of recurrent events in problem-solving items. *Applied Psychological Measurement*, 42, 478–498. <https://doi.org/10.1177/0146621617748325>
- Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education*, 26, 196–212. <https://doi.org/10.1108/QAE-07-2017-0038>
- Yavuz, H. C. (2019). The effects of log data on students' performance. *Journal of Measurement and Evaluation in Education and Psychology*, 10, 378–390.

- Zehner, F., Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment Modeling*, 62(1), 85–105. <https://doi.org/10.25656/01:20354>
- Zenisky, A. L., & Baldwin, P. (2006). Using item response time data in test development and validation: Research with beginning computer users. *Center for educational assessment report* 593.
- Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15, 337–362. https://doi.org/10.1207/S15324818AME1504_02
- Zhan, P., Jiao, H., & Liao, D. (2018). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology*, 71, 262–286. <https://doi.org/10.1111/bmsp.12114>
- Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? an analysis of writing processes. *Educational Measurement: Issues and Practice*, 38, 14–26. <https://doi.org/10.1111/emip.12249>
- Zhang, M., & Deane, P. (2015). Process features in writing: Internal structure and incremental value over product features. *ETS Research Report Series*, 2015, 1–12.
- Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In L. A. van der Ark, D. M. Bolt, W. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 299–314). Springer.
- Zhang, M., Zhu, M., Deane, P., & Guo, H. (2017a). *Identifying and comparing writing process patterns using keystroke logs* (pp. 367–381). The Annual Meeting of the Psychometric Society. Springer.
- Zhang, M., Zou, D., Wu, A. D., Deane, P., & Li, C. (2017b). An investigation of writing processes employed in scenario-based assessment. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 321–339). Springer.
- Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16, S8. <https://doi.org/10.1186/1471-2105-16-S13-S8>
- Zhu, M., Shu, Z., & von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement*, 53, 190–211. <https://doi.org/10.1111/jedm.12107>
- Zhu, M., Zhang, M., & Deane, P. (2019). Analysis of keystroke sequences in writing logs. *ETS Research Report Series*, 2019, 1–16.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.