From byproduct to design factor: on validating the interpretation of process indicators based on log data

Goldhammer, Frank; Hahnel, Carolin; Kroehne, Ulf; Zehner, Fabian

Large-scale Assessments in Education

REVIEW Open Access

From byproduct to design factor: on validating the interpretation of process indicators based on log data

Frank Goldhammer[1,2*], Carolin Hahnel[1,2], Ulf Kroehne[1] and Fabian Zehner[1,2]

*Correspondence: Goldhammer@dipf.de
[1] DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt/Main, Germany
Full list of author information is available at the end of the article

Abstract

International large-scale assessments such as PISA or PIAAC have started to provide public or scientific use files for log data; that is, events, event-related attributes and timestamps of test-takers' interactions with the assessment system. Log data and the process indicators derived from it can be used for many purposes. However, the intended uses and interpretations of process indicators require validation, which here means a theoretical and/or empirical justification that inferences about (latent) attributes of the test-taker's work process are valid. This article reviews and synthesizes measurement concepts from various areas, including the standard assessment paradigm, the continuous assessment approach, the evidence-centered design (ECD) framework, and test validation. Based on this synthesis, we address the questions of how to ensure the valid interpretation of process indicators by means of an evidence-centered design of the task situation, and how to empirically challenge the intended interpretation of process indicators by developing and implementing correlational and/or experimental validation strategies. For this purpose, we explicate the process of reasoning from log data to low-level features and process indicators as the outcome of evidence identification. In this process, contextualizing information from log data is essential in order to reduce interpretative ambiguities regarding the derived process indicators. Finally, we show that empirical validation strategies can be adapted from classical approaches investigating the nomothetic span and construct representation. Two worked examples illustrate possible validation strategies for the design phase of measurements and their empirical evaluation.

Keywords: Log data, Low-level feature, Process indicator, Cognitive assessment, Evidence-centered design, Validation strategies

Typically, in a cognitive assessment, a test-taker completes a series of test items designed beforehand, the work product obtained from each item is scored, and the item scores are aggregated in some way to form a final test score, which is used to infer the test-taker's knowledge, skill, or another cognitive attribute (Mislevy et al., 2003; National Research Council, 2001). This paradigm has been applied in many international large-scale assessments, such as the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies

Springer Open

(PIAAC) by the Organisation for Economic Co-operation and Development (OECD), or the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) by the International Association for the Evaluation of Educational Achievement (IEA). However, the shift from paper-based to computer-based assessment modes in these large-scale assessments (i.e., PIAAC 2012, PISA 2015, TIMSS 2019, and PIRLS 2021) provides an opportunity to assess more than just the work product of each item. The work process itself becomes observable in the form of log data (sometimes also referred to as telemetry data) including all the events, event-related attributes, and timestamps of test-takers' interactions with the assessment system. In light of these new opportunities, log data are becoming more and more frequently available to researchers and other data users in public or scientific use files (for PIAAC see Goldhammer et al., 2020; OECD, 2019).

Knowledge about test-takers' work process may be valuable in many ways from a measurement point of view (for an overview, see Goldhammer et al., 2020). For instance, information derived from log data can be used to capture new process-related constructs (Greiff et al., 2016), support data quality control (Wise, 2017; Yamamoto & Lennon, 2018), increase measurement precision (Klein Entink et al., 2009), validate test score interpretations (Ercikan & Pellegrino, 2017; Li et al., 2017), optimize the test design (van der Linden, 2008), and intervene in the course of test administration (Wise et al., 2019). Many of these uses imply inferring latent (e.g., cognitive or motivational) attributes of the work process from log data (but not all, e.g., increasing measurement precision is simply about exploiting empirical relations). The main focus of this article is that these inferences need to be justified through validation. Both theoretical and empirical evidence is required to ensure that the respective construct interpretation is valid (Goldhammer & Zehner, 2017).

In the following sections, we will review and synthesize concepts from various areas of measurement research. This synthesis provides a novel conceptual underpinning for deriving process indicators from log data and identifies approaches to ensuring their valid interpretation. We will first propose conceptualizing the use of log data from cognitive assessments as a fusion of the standard assessment paradigm with the so-called continuous assessment approach. Afterwards, we will use the evidence-centered design (ECD) framework (Mislevy et al., 2003) to revisit the process of evidentiary reasoning with log data by taking both a theory-based and a data-driven perspective. Most importantly, ECD incorporates the development of the validity argument into the design of the assessment, which in our case is based on log data. Next, we turn our focus to how to empirically challenge the intended interpretation of indicators derived from log data by developing and implementing correlational and/or experimental validation strategies. This will be illustrated with two empirical examples from previous research and finally extended with a discussion and concluding remarks.

Types of assessment

Each type of cognitive assessment can be understood as a process of reasoning from observed evidence (e.g., Scalise, 2012). According to the notion of the assessment triangle (National Research Council, 2001), the key elements of cognitive assessments are a theory or a model of students' *cognition* (e.g., knowledge, skills, competencies), assumptions about what kind of *observations* provide information about these constructs, and an *interpretation* that makes sense of the observational evidence in terms of the target construct. Different types of assessments incorporate these key elements in different ways, as described below.

Standard assessment paradigm

Typical examples of the *standard assessment paradigm* (Mislevy et al., 2012) in educational measurement are competence tests as administered in PISA or PIRLS. Such instruments include a set of pre-defined items, typically consisting of instructions, a question, a stimulus, and a response field, which may be outside or inside the stimulus (an important exception to this structure are units encompassing multiple items related to a single stimulus). The items are designed to obtain locally independent measures; that is, when keeping the construct to be measured constant, the item response variables should no longer be correlated. Thus, the intention is to observe discrete item-by-item responses that do not depend on responses to other items. The observational evidence is typically based on the final work product created when completing an item (e.g., a response selection, a short text). At the individual level, the data obtained from the scored work products are quite sparse and coarse-grained. Assessments following the standard paradigm are obtrusive in the sense that the test-taker has to invest extra time and effort in order to take a test.

Continuous or ongoing assessment

Important examples of the *continuous* or *ongoing assessment* approach (DiCerbo et al., 2016; Mislevy et al., 2012) are game-based assessments and simulation-based assessments. As an example of game-based assessment, *Jackson City* (Mislevy et al., 2014) requires the player to achieve multiple (conflicting) goals by managing activities in a simulated city with complex possibilities for manipulation. The major goals are maximizing economic growth while minimizing pollution. The outcome of the assessment is a measure of the player's system thinking ability (i.e., ability to analyze and control complex relationships). Evidence is collected continuously by evaluating sequences of interactions that take the situational context into account (e.g., shutting down an old power station and building a new environment-friendly one). This type of assessment does not comprise traditional items, but a predefined activity space in which behavior is observed and evaluated continuously. Thus, evidence for the target construct(s) is gathered unobtrusively over time by continuously extracting behavioral indicators from log data under consideration of the context (i.e., with both the player's past behavior and self-dynamics of the game affecting how to interpret the player's behavior in the current situation). The opportunities for obtaining observational evidence are determined not only by the game developer (as in the traditional assessment paradigm), but also by the player, who actively decides which situations to explore. The data can be very rich (at the individual level) and fine-grained. Unlike standard assessments, this type of assessment is less obtrusive or even completely unobtrusive, since data for assessment purposes are collected while the assessed person plays a game or learns in a digital environment (*stealth assessment*, Shute, 2015).

Continuous assessment within cognitive items

Cognitive assessments, including interactive items where the stimulus is manipulated by the test-taker, can be understood as a blend of the standard assessment paradigm and the continuous assessment approach. An example would be a science item that presents a simulated experiment to assess (procedural) knowledge about experimental strategies for inferring rules. What is preserved from the standard assessment paradigm are predefined items. Within each item, however, the provided evidence takes the form of not only the final work product, but also the test-taker's behavior over time capturing attributes of the work process (e.g., presence or absence of a solution strategy). Since such items are less open-ended than in games, for instance, the evidence-generating situations encountered by test-takers are much more comparable and standardized. Continuous assessment within items can be incorporated into the scoring rules for the final work product (e.g., to penalize less efficient strategies). Within items, the extraction of indicators reflecting the work process is unobtrusive. If the items have a sufficient degree of interactivity, the data can be very rich and fine-grained at the individual level.

The ECD view on continuous assessment within items

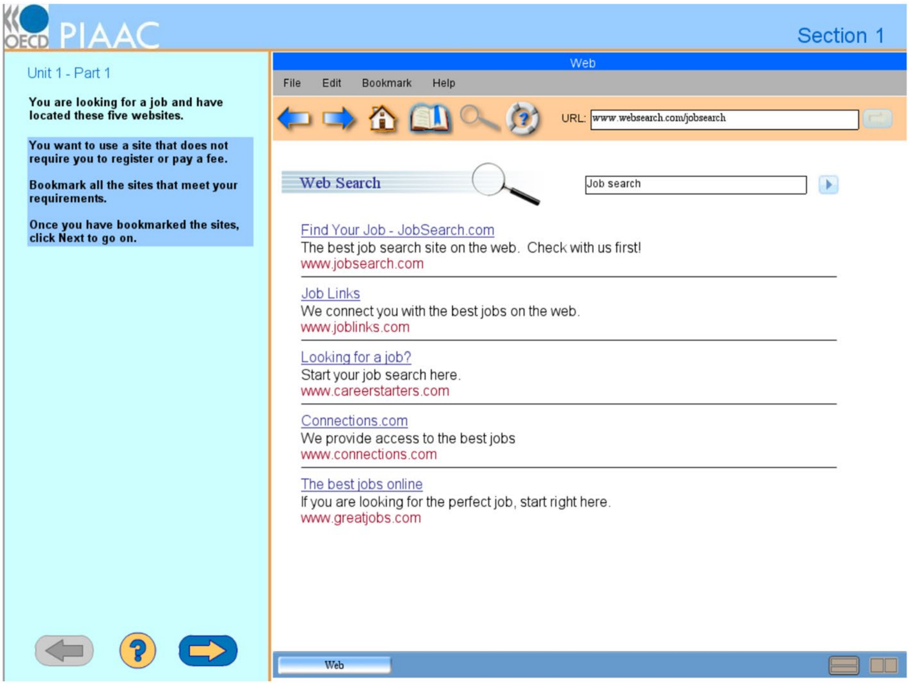
In this section, we present a novel view of the assessment process as reasoning based on log data. It integrates concepts of hierarchical evidentiary reasoning from continuous assessments (Mislevy, 2019) and ECD (Mislevy et al., 2003). Continuous assessment within items provides process indicators that capture latent attributes of the work process. Like product/correctness indicators, process indicators are the result of evidence identification. Thus, they can be conceptualized using the ECD framework, which is a highly flexible approach for designing, producing, and conducting various types of educational assessments. For the present purpose, we focus on the student, evidence, and task models within the ECD framework.

To illustrate the ECD models and validation strategies that will be presented, we begin with the example of the first item from the published PIAAC Job Search Unit (see Fig. 1).[1] The item simulates an interactive web environment and requires test-takers to bookmark all job search-related websites that meet two specific criteria (no registration, no fees); the correct solutions are the websites www.careerstarters.com and www.greatjobs.com.

Student model

In terms of log data, the student (i.e., test-taker, learner, user, or simply person) model includes latent variables representing process-related constructs, such as assumed attributes of the work process. They can be defined in a domain-specific way, such as planning and allocating resources in complex problem solving (e.g., Eichmann et al., 2019; Greiff et al., 2016) or sourcing in reading multiple documents (Hahnel et al., 2019), or more generally, such as test-taking engagement (Goldhammer et al., 2017). As in these examples, the theory-based attributes of the work process can be represented as continuous latent variables or as categorical variables (Greiff et al., 2018). In any case,

[1] https://piaac-logdata.tba-hosting.de/public/problemsolving/JobSearchPart1/pages/jsp1-home.html.



Unit 1 - Part 1

You are looking for a job and have located these five websites.

You want to use a site that does not require you to register or pay a fee.

Bookmark all the sites that meet your requirements.

Once you have bookmarked the sites, click Next to go on.

URL: www.websearch.com/jobsearch

Find Your Job - JobSearch.com
The best job search site on the web. Check with us first!
www.jobsearch.com

Job Links
We connect you with the best jobs on the web.
www.joblinks.com

Looking for a job?
Start your job search here.
www.careerstarters.com

Connections.com
We provide access to the best jobs
www.connections.com

The best jobs online
If you are looking for the perfect job, start right here.
www.greatjobs.com

Fig. 1 PIAAC Job Search unit (Item 1). The screenshot shows the search engine results page presented to the test-taker at the beginning. To solve the item task, test-takers can navigate to the linked pages

they serve to explain systematic individual differences in response behavior. Construct validation requires theoretical and empirical evidence that individual differences in these latent variables can actually be interpreted as determined by differences in the target construct.

In the PIAAC Job Search example (see Fig. 1), the allocation of cognitive resources could be a construct representing an attribute of the work process (Naumann, 2019). Allocation of cognitive resources means that test-takers devote time to processing information that is crucial for solving the task successfully (i.e., spending time examining each website for registration requirements and applicable fees). For difficult tasks that require controlled processing, in particular, the probability of success can be expected to increase if more time is spent on relevant pieces of information (Goldhammer et al., 2014).

Evidence model

The evidence model identifies the kind of observable evidence that is suitable to update the information about the target construct as defined in the student model. For instance, a lack of test-taking engagement may be indicated by responding to an item relatively quickly (i.e., below a certain response time threshold). The evidence model defines how to summarize a person's behavior within an item as an observed variable by applying evidence identification rules. The result of this is a behavioral indicator (i.e., process indicator) capturing an attribute of the work process.

In Fig. 2 (left side), the raw data level with all log events is connected to the construct level by two intermediate inferential steps capturing evidence identification: the

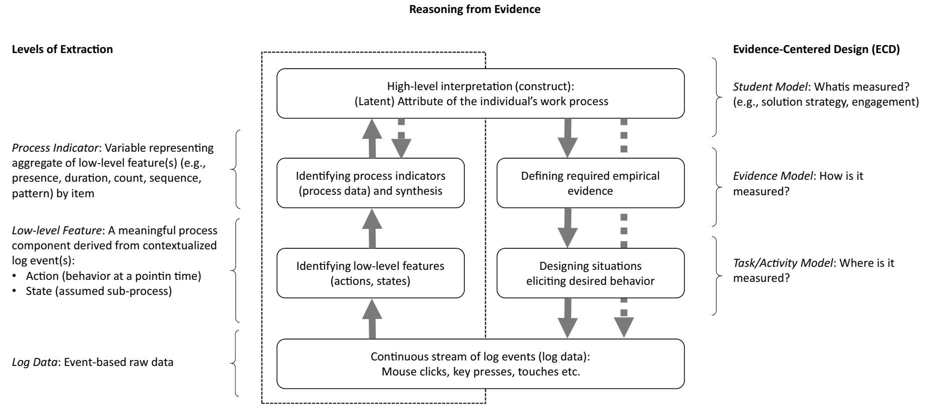

Reasoning from Evidence
Levels of Extraction
Evidence-Centered Design (ECD)
High-level interpretation (construct):
(Latent) Attribute of the individual's work process
Student Model: Whatis measured?
(e.g., solution strategy, engagement)
Process Indicator: Variable representing aggregate of low-level feature(s) (e.g., presence, duration, count, sequence, pattern) by item
Identifying process indicators (process data) and synthesis
Defining required empirical evidence
Evidence Model: How is it measured?
Low-level Feature: A meaningful process component derived from contextualized log event(s):
• Action (behavior at a pointin time)
• State (assumed sub-process)
Identifying low-level features (actions, states)
Designing situations eliciting desired behavior
Task/Activity Model: Where is it measured?
Log Data: Event-based raw data
Continuous stream of log events (log data):
Mouse clicks, key presses, touches etc.

Fig. 2 Assessment as reasoning from evidence derived from log data

identification of low-level features (i.e., action, state) and the identification of process indicators. These two steps are explained in the following sections. The reason for this distinction is that log events that are generated and stored for monitoring and debugging purposes are not per se meaningful and useful from an assessment perspective (Hao & Mislevy, 2018). Moreover, due to the platform-specific format of log data, it can be enormously difficult to compare and reproduce findings if different assessment systems are used (Kroehne & Goldhammer, 2018). Therefore, a more meaningful and generic representation format is desirable, which low-level features are able to provide.

Low-Level features

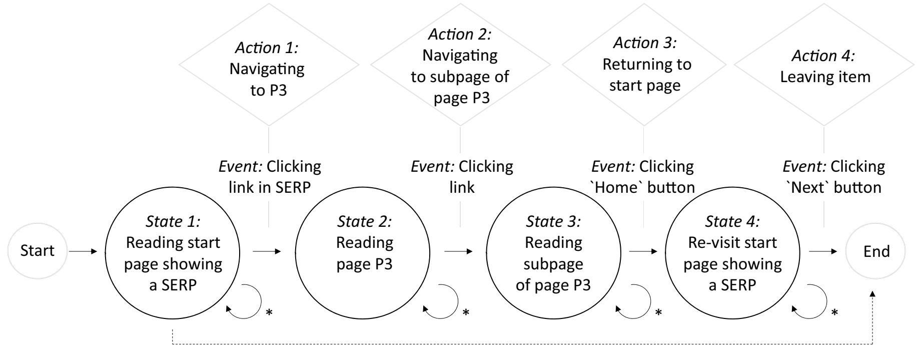
In the first inferential step, the log events are translated into low-level features (Kroehne & Goldhammer, in press). We define a low-level feature as a meaningful process component derived from contextualized log event(s). Hence, low-level features represent a new vocabulary that can be used to represent, compare, and evaluate individual response processes, thus providing a foundation for deriving process indicators for the target attribute. We propose a distinction between two types of low-level features.

First, a low-level feature can represent an *action*, which is defined here as a behavioral act that occurs at a certain point in time. Basically, actions are required to attain a certain goal and are typically controlled by intentions (Ajzen, 1985). In this sense, log events could be translated into *verb clauses* such as "action x at time *t*" (with e.g., x = accessing page P3) representing semantically meaningful (verb-level) user actions in the assessment's activity space (Hao & Mislevy, 2018; Mislevy, 2019). This step can also be referred to as tagging or labelling. As emphasized by Rupp et al., (2012a, 2012b), tags need to be interpretable and meaningful from a substantive perspective (e.g., as actions), which requires defining an appropriate level of granularity. It is important to note that a test-taker's action as identified by log event(s) is also determined by the respective situational context (*contextual dependency of log events*, Kroehne & Goldhammer, in press). That is, the same behavior (e.g., pressing a certain button) might represent different actions depending on the test-taker's past actions and the current situational context. Thus, evidence identification rules need to consider the context of observed behavior and embed it in past (and future) events in the activity space. This important issue will be further illustrated in Example 2 below.

Second, a low-level feature can represent a *state*, that is, a (theoretically) assumed subprocess or a fraction of the test-taker's response process. A state is identifiable by an initiating and terminating log event. The corresponding clause would be "state y at time *t* for duration *d*" (e.g., with y = reading page P3). In order to extract states, the activity space within each item is conceptualized as a set of distinct states and related transitions that are part of one (or multiple) finite state machine(s) (Kroehne & Goldhammer, 2018; Mislevy et al., 2014). States can serve as the building blocks for reconstructing the behavioral response process, which can then be associated with cognitive processing (e.g., the state of reading the instructions, the state of reading and evaluating a certain page in the stimulus, the state of making a response). The assumed information processing states are defined based on the (theoretical) model of the response process being applied and the research questions being addressed. They are empirically identified by relating transitions between them to observable log events (Kroehne & Goldhammer, 2018). A test-taker's transition from one state to another may be the result of the test-taker's behavior or be initiated by the assessment system itself. With respect to the latter, the system can function as an actor whose actions depend on test-takers' behavior (e.g., virtual agent that provides support if needed) or are independent of it (self-dynamic, e.g., the information presented to the test-taker changes after a certain amount of time regardless of the test-taker's actions). Note that log events identifying the transition between states (as well as log events within states) can represent meaningful actions by the test-taker or the system that are of interest as low-level features.

It is the researcher's decision with what degree of granularity states and actions should be defined given the availability of identifying log events. If more coarse-grained state definitions are applied, actions can also be defined to occur within states (i.e., not all actions are necessarily transitions between states). If fine-grained states are used, for instance, to provide detailed contextual information regarding the stream of events, actions can be identified by log events being contextualized by these states. A sequence of actions can also serve as a proxy for the test-taking process when no differentiation into states is possible or necessary. In our example, within the state 'reading page P3' we might also observe scroll events that do not allow for defining states identifying exactly which parts of the text are being read, but that could be used to identify specific actions. Note that our definition of an action does not assume that an action has a meaningful temporal duration. In the example above, the action of 'accessing page P3' by clicking on a link has the temporal duration of how long the mouse is pressed down (identified by the log events of pressing and releasing the mouse button). However, this duration is not considered relevant. In contrast, the state of 'reading page P3' does exhibit a temporal extension that is presumably intentionally controlled and meaningful.

Referring once again to the PIAAC Job Search task, Fig. 3 shows a simple example illustrating how a sequence of actions and inferred states capture the process of completing an item. As described above, actions and states serve to decompose the test-taking process into low-level features that are theoretically meaningful. Accordingly, the test-taker in the example begins by reading the start page (State 1), then clicks on a hit from the search engine results page to visit the corresponding website (Action 1), reads


Action 1: Navigating to P3
Action 2: Navigating to subpage of page P3
Action 3: Returning to start page
Action 4: Leaving item
Event: Clicking link in SERP
Event: Clicking link
Event: Clicking `Home` button
Event: Clicking `Next` button
Start
State 1: Reading start page showing a SERP
State 2: Reading page P3
State 3: Reading subpage of page P3
State 4: Re-visit start page showing a SERP
End
*

······ One possible alternative process, leaving the item without interaction with the stimulus.

* Additional log events not considered in this example.

Fig. 3 Process of completing an item represented as a sequence of actions and states (SERP = search engine results page)

information on this website (State 2), clicks on another link to move on to a subpage (Action 2), reads information on this page (State 3), and then clicks on the browser home button to return to the start page (Action 3). After reading the start page again (State 4), the test-taker clicks the next button to leave the item (Action 4) without providing a response by bookmarking a website. Note that Fig. 3 reflects only one possible sequence in terms of actions and states.

In the PIAAC Job Search example (see Fig. 1), the time spent on relevant pages could provide evidence for the allocation of cognitive resources. Relevant pages can be defined as pages containing information needed to solve the task or that need to be visited to get there. A test-taker's log file capturing the continuous stream of log events may have the following two successive entries:

```
<taoEvent Name="stimulus" Type="TEXTLINK" Time="164959">id=u10a_default_txt
15|*$href=unit10page14|*$target=_self</taoEvent>
<taoEvent Name="stimulus" Type="TOOLBAR" Time="166984">id=toolbar_home_
btn</taoEvent>
```

Using knowledge about the item and assessment system, these event-based raw data can be translated into low-level features, that is, two time-stamped actions: the action 'navigate from the search engine results page to the website www.greatjobs.com' at time 164,959 and the action 'navigate from the website www.greatjobs.com to the search engine results page by clicking on the home button' at time 166,984. The state 'visiting the relevant page www.greatjobs.com' at time 164,959 for a duration of 2025 ms can also be derived from these events.

Process indicators

In the second inferential step, the obtained low-level features are summarized into item-level indicators of the target attribute of the work process by applying evidence identification rules. This could simply be a count measure representing whether or how often a certain action or state has been observed, or a more complex process indicator representing a certain solution strategy in the form a sequence or pattern of actions and/or states (e.g., control-of-variables strategy in a science item presenting a simulated experiment). Note that low-level features and related patterns may be specific to individual items, while the process indicators derived from them are defined across items (e.g., the same solution strategy could be captured by different combinations of low-level features in different items).

In the standard assessment paradigm, the application of evidence identification rules means scoring the work product created by the test-taker. However, when "scoring" the work process (i.e., creating a process indicator), evidence identification refers instead to identifying the presence or absence of certain low-level features (i.e., actions or states) or certain patterns of low-level features in the continuous stream of log events elicited during the item completion process. Accordingly, evidence identification in the traditional sense is scoring a time-bounded work product, whereas in continuous assessment, it refers to feature extraction and aggregation based on continuous behavior over time (Behrens & DiCerbo, 2014). Note that a time-bounded work product could also be extracted from log data as long as all (final) response-related events are available (Kroehne & Goldhammer, 2018).

As in the case of a latent variable, construct validation requires justifying that the observed individual differences in a process indicator can (unambiguously) be related to individual differences in the construct of interest. If the process indicator captures a theoretically defined construct, this theory should (ideally) be used to determine the evidence needed and the kind of identification rule that would be appropriate. The development of evidence identification rules and the extraction of evidence can be systematically planned using the in-task assessment framework (I-TAF), which proposes a cognitively enhanced ontology linking (low-level) features to constructs (Kerr et al., 2016). An overview of tools supporting the extraction of features is provided by Kroehne and Goldhammer (in press) and includes the R package logFSM (Kroehne, 2021) for analyzing log data using finite-state machines.

In our PIAAC Job Search example, the item-level process indicator reflecting the construct 'allocation of cognitive resources' would be the time spent on relevant pages, obtained as the aggregated (e.g., sum, average) duration of all states capturing visiting a relevant page. Construct-irrelevant variance would be induced if a test-taker visits a relevant page and spends some time on the page without being engaged in solving the task (e.g., due to distracting thoughts).

In order to obtain a reliable measure that comprehensively captures the target construct, item-level process indicators need to be synthesized into a test-level indicator. This can be accomplished with statistical models such as standard psychometric measurement models or Bayesian networks (for a review of methods for performance data from simulation-based assessments, see de Klerk et al., 2015). A challenge here is to fully capture the dependency structure of process (and product) indicators within and between items, which can be more complex than in the standard assessment paradigm, which only considers the final work product of each item (for a discussion of measurement models see, e.g., Levy, 2020).

Task/activity model

Design of activity spaces

The task model is about designing situations in a way that generates the evidence needed to make inferences about the target construct. With respect to continuous assessments within items, this refers to designing activity spaces within items that can elicit the desired behavioral evidence in the form of a sequence of actions and states during test-takers' interaction with the task environment. Following Behrens and DiCerbo (2014), standard assessment and continuous assessment can be differentiated by a shift from the item paradigm to the activity paradigm. Items pose questions with responses as output, this output is scored, and the obtained variable is interpreted as a correctness or product indicator of a construct (e.g., competence). In short, the information provided by items is focused. Activity spaces request or invite interaction with extracted low-level features as output representing actions or states. This output is "scored" (i.e., aggregated or summarized), and the obtained variable is interpreted as a process indicator of a construct (i.e., attribute of the work process). The information provided in continuous assessments can be rich, allowing multiple aspects of the work process to be described.

The valid interpretation of process indicators depends on a careful and clear definition of how the target attribute, empirical evidence (behavioral low-level features and derived process indicators), and activity space that can elicit the desired behavior are linked to each other. Specifically, the task must be designed in a way that generates an activity space within items so that to-be-inferred attributes of the work process can be linked to behavior (Goldhammer & Zehner, 2017). This also has implications for the system design, as user (and system) events need to be stored correctly and completely. The granularity and completeness of event logging depend on the (low-level) features to be extracted (for completeness conditions, see Kroehne & Goldhammer, 2018; Oranje et al., 2017).

In the PIAAC Job Search example, it is possible to identify the state 'allocating cognitive resources to relevant pages' by means of log events due to the design of a search engine results page with linked and clickable websites providing relevant or irrelevant information. In contrast, in a task presenting only a search engine results page and requiring test-takers to select a suitable source based on the available information snippets (page title, excerpt, URL), it would not be possible to determine the allocation of time to different snippets by means of log events (and other types of process data would instead by required, such as data on eye movements).

Sampling observations

Activity spaces within items need to be designed to ensure that the actually observed behaviors are a representative sample of the universe of possible observations (generalization inference; Kane, 2013). In general, the generalization inference is justified by including a representative sampling of items (e.g., in terms of context, structure, complexity) in a test.

In continuous assessments involving rich simulations in a large activity space (e.g., game-based assessment) rather than traditional items, the task situations encountered might differ between individuals, making it difficult to ensure a representative sampling of observations. This is because, in continuous assessments, individuals have some control over the task situation through their choice of actions. For instance, individuals may not have the opportunity to demonstrate some behaviors because they do not reach a certain sub-space of the task environment.

This may also be a problem for continuous assessments within items, although it is less severe given the preservation of the 'item' structure and less open-ended character of the items. To avoid having the breadth of the activity space compromise the sampling of observational evidence, a focus on identifying salient features in recurring situations is recommended (Mislevy et al., 2012). Another strategy is to align task situations across individuals by introducing rescue/convergence points. If test-takers are on the wrong track or lost in the activity space, they can be re-located—for instance, by providing required information that enables them to continue and to demonstrate the behavior needed for evidence identification (e.g., "rescue" agents in the collaborative problem solving assessment in PISA 2015, OECD, 2017). This strategy also helps to control the testing time and mitigate the problem of local dependencies between indicators. The latter means reducing or eliminating the dependency of a test-taker's performance on the current part of the problem on his or her performance (i.e., success vs. failure) on a previous part of the problem. Note that if failing to reach a certain task situation is due to the target construct (e.g., lack of skill or knowledge), this observation could serve as evidence for this construct.

The PIAAC Job Search example item is designed in a way that enables a comparable sampling of evidence for the process indicator (i.e., time on relevant pages) across test-takers. For this task model, visiting pages linked to the search engine results page is a salient feature, and most if not all test-takers can be expected to demonstrate such behavior. An exception may be test-takers without knowledge of how to navigate between linked pages in web environments.

Theory- and data-driven reasoning from log data

Figure 2 shows continuous assessment as a reasoning process from log data to higher-level interpretations (construct level). This reasoning process may include data-driven and theory-driven elements to varying degrees (see the concept of "computational psychometrics" integrating psychometrics and data mining, e.g., Drachsler & Goldhammer, 2020; Rupp et al., 2012; von Davier, 2017).

The data-driven construction of process indicators (see left side of Fig. 2) typically relies on data-mining techniques to explore patterns or regularities in low-level features derived from log data. It can be unsupervised, such as when low-level features obtained from an item are clustered to learn about underlying structures in the data. An individual's membership to a certain cluster is then interpreted in terms of an attribute of the work process (e.g., as an indication of a certain solution strategy, Eichmann et al., 2020; Ulitzsch et al., 2021), which requires a theoretical model in the background to enable the derivation of a construct explaining behavioral differences (see dashed arrow on the left side of Fig. 2). Supervised approaches use (low-level) features to predict continuous outcomes or categorical outcomes, with a focus on learning the mapping function between the input and outcome variables in order to predict outcomes based on new input data. Continuous output variables may be valid standardized measures or expert ratings (e.g., Margolis & Clauser, 2006). If the predictions are successful, the low-level features can serve as an unobtrusive alternative to traditional assessment. A major categorical outcome is task success (e.g., Han et al., 2019; He & Von Davier, 2016). Predicting it via low-level features can shed light on behavioral—and related cognitive—processes enabling or hampering successful task completion. In the case of transparent machine learning approaches (e.g., decision trees), the prediction model (i.e., how features are selected, combined, and weighted) may be useful for generating interpretations and hypotheses about underlying differences in strategies and misconceptions. This allows process indicators to be derived in a data-driven and theory-driven way. Finally, it is important to examine the trained evidence identification rule with data that was not included in building the prediction model.

The selection of the either supervised or unsupervised approach is mainly determined by the availability of external optimization criteria and established theories that can be adopted. On the one hand, supervised approaches require at least one criterion to be predicted. In turn, this criterion to be optimized must come with some theoretically and/or empirically supported validity argument (Kane, 2013). The impact of established theories after an appropriate and accurate prediction model has been built through supervised optimization is limited. However, transparent statistical modelling can enhance the alignment between the new prediction model and established theories and generate hypotheses. In the case of a lack of fit to established theories or opaque machine learning, the prediction model's generalizability to independent samples must be demonstrated. On the other hand, unsupervised approaches such as clustering require theoretical underpinnings or domain-expert knowledge to map the resulting model entities to the desired assessment categories. Accordingly, the selection of either approach implies a range of available and required validation steps.

The fully theory-driven construction of process indicators (see right side of Fig. 2) starts with the assessment framework defining the attribute of the work process to be measured. A construct may be defined based on a cognitive model of information processing or some other theoretical rationale providing information on what kind of evidence is needed and how to design task situations (i.e., activity spaces) and the assessment system to elicit and observe the desired actions and states needed for evidence identification (e.g., Abele & von Davier, 2019; Hahnel et al., 2019). Evidence identification is based on theoretical assumptions about low-level features (i.e., actions and states) that capture the test-taker's information processing when interacting with the item, and how these low-level features can be identified through log events. Then, rule-based functions are applied to log events to extract low-level features (i.e., actions and states), which are then aggregated to obtain the process indicators.

The dashed arrow on the right side of Fig. 2 from the construct level to the raw log data level illustrates that not all assessments will be designed from the beginning with the goal of using log data. Log data available from such assessments can be an interesting source for the theory-driven extraction of low-level features and process indicators. The usefulness of these log data, however, may be limited because they are not sufficient to identify the desired actions and states. This may be due to inadequate interactivity in the task/activity model or simply for technical reasons (e.g., relevant log events were not stored by the assessment system). Note that in this situation the validation strategy for process indicators lacks theoretical a-priori arguments that have driven the item design.

Argument-Based validation

Indicator-Based inferences

Following Kane (2001, 2013), central inferences when interpreting indicators are scoring or evaluation, generalization, explanation, extrapolation, and decision making. In this section, we will focus on the explanation inference, which is essential when interpreting process indicators in terms of theory-based attributes of the work process. This inference, however, also requires that the scoring inference (see the previous section *Evidence Model*) and the generalization inference (see previous section *Sampling Observations*) are justifiable. The scoring or evaluation inference is made when applying evidence identification rules to behavioral observations. What is obtained is an observed indicator variable (product or process indicator). This first inference is based on the assumption that the evidence identification rules appropriately extract the targeted behavioral features. The generalization inference was mentioned in the previous section about sampling observations. It refers to generalizing from an individual's behavior in a specific assessment to behavior in similar tasks under similar conditions. This requires us to be able to observe a representative sample of an individual's behavior from the universe of possible observations as defined by the construct to be assessed.

Following the Standards for Educational and Psychological Testing, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. […] Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (AERA et al., 2014, p. 4; see also Messick, 1989). These concepts of validity and validation apply to any indicator-based inferences, regardless of whether product/correctness or process indicators are used. Thus, inferring latent (e.g., cognitive) attributes of the work process from indicators needs to be justifiable (Goldhammer & Zehner, 2017). Validation refers to the process of developing and evaluating arguments speaking for and against a certain interpretation and use of an indicator (Kane, 2013). This requires specifying the interpretation and use of the indicator as well as explicating related assumptions, including the line of reasoning from behavior to the intended inference. Finally, the argument is evaluated both conceptually and empirically.

Explanation inference and related empirical sources of validity evidence

The explanation inference means that individual differences in the process indicators at the item level and their aggregation to an indicator at the test level are (causally) determined by differences in the (theoretical) construct which the indicator is intended to measure. Such a theory-based interpretation requires a theory that defines the construct (i.e., attribute of the work process, e.g., solution strategy) and describes how the construct explains behavioral differences captured by the process indicator. In general, empirical validity evidence is provided when there is empirical support for theory-based predictions about relationships between observable variables (including the respective process indicator).

Threats to the construct interpretation of indicators are construct-irrelevant variance and construct underrepresentation (Huff & Sireci, 2001; Messick, 1989). Construct-irrelevant variance means that other sources of variance apart from the target construct affect the observed behavioral differences captured in the indicator. That is, reasoning from the observations to the target construct becomes ambiguous. Construct underrepresentation means that the indicator is too narrow, as relevant aspects of the construct are missing or not represented by the indicator. That is, reasoning from the observations may be flawed as the target construct is not fully captured.

The empirical validation of the construct interpretation (i.e., the explanation inference) of process indicators can be conducted by adopting approaches commonly used for correctness indicators. In the following, we will discuss the construct representation and the nomothetic span approach, focusing on differences at the item level and person level (Embretson, 1983).

Construct representation

"Construct representation is concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores" (Embretson, 1983, p. 179). Although this definition refers to item responses (i.e., product indicators), the underlying notion can be generalized to process indicators capturing an attribute of the work process. Accordingly, task characteristics that theoretically evoke the target attribute are determined. For instance, a certain property of the stimulus may be expected to increase the probability of evoking a particular cognitive process such as applying a certain solution strategy. The task characteristic (or property of the stimulus) is then related to observable item-level process indicators measuring the target attribute. If items with this particular task characteristic are actually more likely to elicit the respective sequence of actions or states representing the target cognitive process, then the process indicators can be interpreted as determined by this latent attribute of the work process. Statistical models, such as the family of explanatory item-response models, are available to easily implement this validation approach (e.g., lltm + e, Janssen et al., 2004).

Traditional experimental designs are also suitable to collect validity evidence for the construct interpretation of process indicators. If the theory defining the construct suggests factors that are clearly expected to influence the corresponding attribute of the work process, the assumed effects can be tested experimentally to provide support for the (causal) explanation inference regarding the process indicator. For instance, providing incentives in a low-stakes assessment (e.g., monetary reward, Braun et al., 2011) or turning a low-stakes into a high-stakes assessment (e.g., by manipulating the instructions) can be expected to increase test-taking engagement, which should in turn be reflected in the process indicator of test-taking engagement (e.g., response time effort, RTE, see Wise & Kong, 2005).

Let us return to the PIAAC Job Search example from above to illustrate the construct representation approach. The extent to which cognitive resources and time need to be allocated may theoretically depend on how and where the information about whether the two search criteria are met is presented. Recall that these search criteria are no fee and not needing to register. This information could be presented in a more easily accessible (list of bullet points) or less easily accessible manner (embedded in a paragraph of text), as well as closer to or further from the start page (e.g., the node distance from the search engine results page to the respective critical information in the hyperlink graph is 2 clicks for www.careerstarters.com and 1 click for www.greatjobs.com). Likewise, the critical information could be presented jointly on one (as in the example item) or separately on two different pages. These stimulus characteristics can be assumed to affect the need to allocate cognitive resources within the item. For instance, if more clicks are required to reach the critical information in an otherwise comparable item, and more time is spent on relevant pages in this item, this would provide some evidence for the construct interpretation of the process indicator as reflecting the allocation of cognitive resources.

Nomothetic span

"Nomothetic span is concerned with the network of relationships of a test score with other variables" (Embretson, 1983, p. 179). Although this definition originally refers to traditional test scores, it can be generalized to indicators derived from log data. These other measures can represent the same or a similar construct, providing convergent evidence, or a different construct, providing discriminant evidence.

A major strategy for obtaining convergent evidence is to investigate the relation between the process indicator and a standardized measure ostensibly measuring a similar construct. For instance, an assessment of self-regulation in a digital learning environment based on log data could be related to a standardized measure of self-regulated learning (for help seeking, see e.g., Aleven et al., 2010).

Another strategy for supporting the interpretation of a process indicator derived from log data is to triangulate it with other methods and data sources obtained from a particular cognitive assessment. Here, the process indicator is related to other measures of the same attribute of the work process, for instance, measures based on think-aloud protocols, eye-tracking or video recordings of the test-taker and the screen (Maddox, 2017). These additional measures should already be validated to some extent and interpretable in terms of the target attribute. This approach can be extended into a multitrait-multimethod analysis by considering different attributes of the work process measured with different methods (Campbell & Fiske, 1959).

Another approach is to relate individual differences in a process indicator to differences in the product or correctness indicator. If there exists a cognitive process model or some conceptual rationale for certain hypotheses about the relation between process indicators and product indicators, the assumed association can be tested empirically (for a validation of indicators of test-taking engagement, see, e.g., Lee & Jia, 2014, and the example below).

Finally, process indicators can be related to group variables to test whether group membership is (theoretically) related to certain attributes of the work process. For instance, experts and novices can typically be expected to differ in their solution strategies (for an example from game-based assessment, see DiCerbo et al., 2011).

In the PIAAC Job Search example, the nomothetic span approach could be applied by correlating individual differences in the allocation of cognitive resources with theoretically related variables, such as comprehension skills, strategy knowledge, and motivation (Naumann, 2019), and testing whether the expected relational pattern is found.

Validation examples

In order to demonstrate the argument-based validation of process indicators, we selected two examples from previous research differing in their use of a process indicator, the kind of process indicator, and the employed validation strategy. In the first example (from Goldhammer et al., 2016), process indicators of test-taking engagement are used for quality assurance in a large-scale assessment (PIAAC). They represent generic indicators based on the total time on task, and the nomothetic span approach is employed for validation. In the second example (from Hahnel et al., 2019), process indicators of sourcing serve to address substantive research questions in the domain of multiple document reading. The process indicators are highly domain-specific and were created by contextualizing selected log events. The construct representation and the nomothetic span approach were used for validation.

Example 1: test-taking engagement

Low test-taking engagement refers to the phenomenon that test-takers do not make an effort to show what they know or can do, but respond quickly and arbitrarily (e.g., Wise & DeMars, 2005). Negative consequences are manifold. Test scores may underestimate the true proficiency level, construct-irrelevant variance is introduced, and the validity of inferences based on test scores may be compromised (Haladyna & Downing, 2004; Kong et al., 2007). Consequently, it is important to detect disengaged responses and take this information into account in scoring and data analysis. Evidence models for the assessment of test-taking engagement are often based on observed response times. The idea is straightforward and requires defining a response time threshold separating disengaged response behavior (i.e., fast [non-]responses, rapid guessing) from engaged response behavior (i.e., taking the time to complete the item). Constant thresholds have been proposed, such as a three-second rule (Kong et al., 2007), as have item-specific thresholds (e.g., Lee & Jia, 2014; Wise & Kong, 2005). One way to define item-specific thresholds is to visually identify the gap in the bimodal response time distribution that separates disengaged and engaged responders (visual inspection, VI). An alternative method is to compute the proportion correct conditional on response time and define the threshold as the time when the proportion correct (P+) exceeds chance level, which was usually 0% in PIAAC (P + >0% method) (Goldhammer et al., 2016).

The intended interpretation of the process indicator was that the differences captured in the indicator are determined by the test-taking engagement construct. For validation, Goldhammer et al. (2016) proposed a set of testable assumptions (see also Lee & Jia, 2014). For the VI method, they provided support for the evaluation inference by showing that independent human raters were able to apply the evidence identification rules consistently. Moreover, supporting the construct interpretation, they related the process indicators to product indicators by comparing the proportion correct for engaged and disengaged responding. For engaged responding, they expected the probability of obtaining a correct response to be much higher than chance level, whereas for disengaged responding, it should be only chance level. Note that for indicators based on the P + >0% method, the latter property was already part of the design of the indicator. Table 1 shows the aggregated results by domain and method. All methods worked quite well. The greatest difference was obtained for the P + >0% method, that is, the low proportion correct of 0% for disengaged responding did not occur at the expense of a higher rate of false negatives for engaged responding. As another source of validity evidence, Goldhammer et al. (2016) correlated the score group (proficiency) with proportion

Table 1 Average proportion correct for engaged and disengaged response behavior in PIAAC 2012 (round 1) by domain and method

	Method	Proportion correct—Engaged	Proportion correct—Disengaged	Difference
Literacy	5000	.55	.02	.53
	3000	.55	.01	.54
	VI	.56	.02	.54
	P + > 0%	.56	.00	.56
Numeracy	5000	.64	.09	.55
	3000	.63	.04	.59
	VI	.63	.07	.56
	P + > 0%	.63	.00	.63
Problem solving	5000	.40	.00	.40
	3000	.40	.00	.40
	VI	.43	.01	.42
	P + > 0%	.43	.00	.43

The table is from Goldhammer et al., (2016, p. 19)

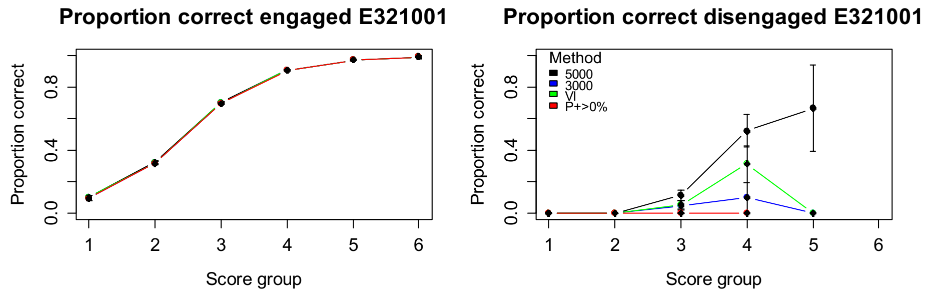


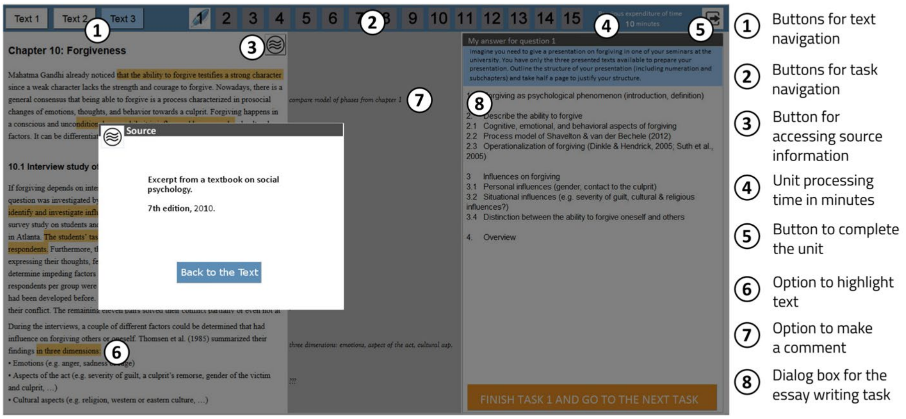
Fig. 4 Relationship between score group and proportion correct for literacy (selected item) (figure from Goldhammer et al., 2016, p. 24)

correct by item, expecting a positive relation for engaged responding and no relation for disengaged responding. Figure 4 shows for a selected PIAAC item that this assumption was supported for all methods except using a constant threshold of five seconds. Obviously, a fair number of false positives were included in the group of disengaged responses, making the relation between score group and proportion correct positive.

Example 2: sourcing in reading

Multiple document comprehension (MDC) is a reader's competence in constructing an integrated representation of a certain topic using textual information from different sources. The MDC test by Schoor et al. (2020) was designed for continuous assessment within MDC items, with the goal of inferring *sourcing* as an important attribute of the work process. Sourcing is defined as the reader's consideration of the origin and intention of a document.

In the task/activity model for the assessment of sourcing, Schoor et al. (2020) designed the activity space within MDC items such that sourcing can be linked to observed behavior (i.e., a state that can be interpreted as inspecting the source information for a text


Text 1
Text 2
Text 3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Previous expenditure of time
10 minutes
Source
Back to the Text
FINISH TASK 1 AND GO TO THE NEXT TASK
1 Buttons for text navigation
2 Buttons for task navigation
3 Button for accessing source information
4 Unit processing time in minutes
5 Button to complete the unit
6 Option to highlight text
7 Option to make a comment
8 Dialog box for the essay writing task

Chapter 10: Forgiveness

Mahatma Gandhi already noticed that the ability to forgive testifies a strong character since a weak character lacks the strength and courage to forgive. Nowadays, there is a general consensus that being able to forgive is a process characterized in prosocial changes of emotions, thoughts, and behavior towards a culprit. Forgiving happens in a conscious and uncondition factors. It can be differentiat

10.1 Interview study of

If forgiving depends on inte question was investigated by identify and investigate infl survey study on students an in Atlanta. The students' tas respondents. Furthermore, t expressing their thoughts, fe determine impeding factors respondents per group were had been developed before. their conflict. The remaining eleven pairs solved their conflict partially or even not at

During the interviews, a couple of different factors could be determined that had influence on forgiving others or oneself. Thomsen et al. (1985) summarized their findings in three dimensions:

- Emotions (e.g. anger, sadness, rage)
- Aspects of the act (e.g. severity of guilt, a culprit's remorse, gender of the victim and culprit, …)
- Cultural aspects (e.g. religion, western or eastern culture, …)

compare model of phases from chapter 1

three dimensions: emotions, aspect of the act, cultural asp.

???

Excerpt from a textbook on social psychology.

7th edition, 2010.

My answer for question 1

Imagine you need to give a presentation on forgiving in one of your seminars at the university. You have only the three presented texts available to prepare your presentation. Outline the structure of your presentation (including numeration and subchapters) and take half a page to justify your structure.

1. Forgiving as psychological phenomenon (introduction, definition)

2. Describe the ability to forgive
2.1 Cognitive, emotional, and behavioral aspects of forgiving
2.2 Process model of Shavelton & van der Bechele (2012)
2.3 Operationalization of forgiving (Dinkle & Hendrick, 2005; Suth et al., 2005)

3 Influences on forgiving
3.1 Personal influences (gender, contact to the culprit)
3.2 Situational influences (e.g. severity of guilt, cultural & religious influences?)
3.4 Distinction between the ability to forgive oneself and others

4. Overview

Fig. 5 Example unit assessing multiple document comprehension. The screenshot shows the dialog box with the source information after the reader has clicked on the source button (3)

could be identified using log events). This was achieved by requiring the reader to click on a button to access the source information for a document (see Fig. 5). Clicking on the source button opened a dialog box presenting the source information (e.g., author, type of document). The box could be closed by clicking the back-to-the-text button. Test-takers were familiarized with this functionality in the tutorial for the MDC test.

Following previous research on sourcing, Hahnel et al. (2019) distinguished three kinds of sourcing depending on purpose: proactive, repeated, and task-related sourcing. The evidence model then defined evidence identification rules for each kind of sourcing. In each case, clicking on the source button defined a sourcing action and a sourcing state (together with clicking the back-to-the-text button). The sourcing state was further contextualized depending on the pattern of past and future states, including temporal information, to identify the three kinds of sourcing. Accessing the source information indicated proactive sourcing if the source was accessed for the first time and within the first 10% of document processing time. Repeated sourcing was indicated if the same source information had been accessed. Finally, task-related sourcing was indicated when the reader switched from the item posing a question to one of the documents and its source immediately (i.e., after spending a maximum 10 s on the document). In the following illustration, we focus on repeated sourcing.

The intended interpretation of the process indicator was that behavioral differences are due to the construct of repeated sourcing. Based on previous research, repeated sourcing is assumed to update memory traces to strengthen mental connections or to help resolve conflicts across multiple documents. Based on prior research on MDC and sourcing, Hahnel et al. (2019) proposed a series of testable assumptions: i) repeated sourcing is positively associated with MDC, but not with final school grades after controlling for MDC; ii) the number of documents, number of conflicts between documents, and number of items that require comprehending source information should induce more repeated sourcing. Thus, the sources of empirical evidence for the construct interpretation of the repeated sourcing indicator were the association with domain-specific competence scores and other measures (nomothetic span), as well as with characteristics of units, as collections of particular documents and items (construct representation).

Validity evidence was obtained by predicting the binary unit-level indicator of *Repeated Sourcing* (0 = source was not accessed or only once; 1 = source was accessed multiple times) with person-level and unit-level variables. The results showed a positive effect of MDC scores and a non-significant effect of final school grades. In terms of unit characteristics, the number of documents and number of conflicts showed positive significant effects, as expected, while the effect of the number of source-related items was not significant. Thus, the findings provide some first validity evidence; however, as also stressed by the authors, further validation is needed.

Discussion

In this paper, we focused on complex interactive items from cognitive (large-scale) assessments following the standard assessment paradigm (see e.g. the PIAAC item in Fig. 1). In this case, product data (i.e., product or correctness indicators) provide information about whether or not the task goal was achieved successfully and typically serve to measure a latent ability or competence construct. In comparison, process data (more precisely, process indicators) can be used to measure a latent attribute of the work process, thus providing information about how the task was approached in order to achieve the task goal (e.g., speed, engagement, strategy use such as sourcing).

Validation requires first specifying the intended interpretation of the indicator (regardless of whether it is defined at the item or test level, whether it is based on product data or process data, and regardless of the type of assessment). From this specification follows the range of sources of evidence that may be suitable to support the respective interpretation and inference. In the present paper, we focused on the explanation inference and construct interpretation of process indicators, respectively, and considered several sources of evidence related to the validity framework of the Standards for Educational and Psychological Testing (AERA et al., 2014). Our discussion of the generalization inference and the sampling of observations referred to *evidence based on test content*. Typically, this kind of evidence is provided by conducting logical or empirical analyses of whether the target construct is fully represented by the test content and the given opportunities to collect required evidence. *Evidence based on relations to other variables* was provided in Example 1 and Example 2, where the nomothetic span approach was used (relation to task success, relations to person characteristics). *Evidence based on the response process* supports the claim that construct differences causally determine differences in the observable indicators. Related evidence can be provided using the construct representation approach; see for instance Example 2, which investigated whether theoretically expected effects of unit characteristics can be found empirically. *Evidence based on internal structure* refers to the extent to which relationships between (item-level) indicators correspond to the target construct. This kind of evidence could be provided by analyzing and testing the expected dimensional structure. For instance, related to Example 1, Goldhammer et al. (2017) investigated whether a uni-dimensional Rasch model fits the process data (item-level test-taking engagement indicators).

Depending on how a construct is defined, both product data and process data may be related to the same latent construct. This is appropriate if process data provides evidence about the respective ability or competence construct over and above the achievement of a correct solution. In this case, process data can be added to the measurement model of the ability or competence construct in the evidence identification or evidence synthesis phase. With respect to evidence identification, process data can enable a more fine-grained (partial credit) scoring of the work product. For instance, if efficiency is part of the competence construct, full credit is given for a correct work product only if the number of actions falls below a certain threshold; otherwise, partial credit is given (e.g., problem solving in PISA 2012; OECD, 2013). Thus, properties of the work process are part of the competence or ability construct (for a related discussion about speed as a nuisance factor or part of the ability construct, see Goldhammer, 2015; van der Linden, 2005). In terms of evidence synthesis, product and process indicators could be used to jointly measure a construct. For instance, De Boeck et al. (2017) argue that the latent variables for capacity and speed-accuracy balance (response cautiousness) are equivalent to the latent variables for speed and ability, but explain response time and response accuracy data differently. Capacity is positively related to both response accuracy and response speed, whereas balance is positively related to response accuracy and negatively related to response speed.

The two empirical validation examples presented in this paper refer to situations where the product data and process data capture different latent constructs. In Example 1, response time is transformed into item-level engagement indicators representing the construct of test-taking engagement, whereas response accuracy captures the respective competence construct. In Example 2, product data (i.e., response accuracy) capture the multiple document comprehension construct, and process data are used to measure sourcing as an attribute of how the multiple documents were read. The presented (construct) validation approaches can be applied similarly even if product and process indicators define a construct jointly. For instance, a straightforward way of validating the 'capacity' (De Boeck et al., 2017) interpretation would be to correlate the capacity variable with a theoretically related variable. Based on the simple view of reading (Hoover & Tunmer, 2018), the accurate and rapid identification of words is assumed to be a limiting factor in comprehension and should exhibit a high correlation with reading comprehension (nomothetic span approach; Goldhammer et al., 2021). Moreover, assuming that word frequency represents a construct-related item characteristic, it could be tested whether word frequency (Gerhand & Barry, 1999) has a positive effect on response accuracy and response speed in a lexical decision task (construct representation approach).

Process indicators can be constructed in different ways based on a given sequence of actions and states (e.g., time measures, count measures, sequential measures). The way they are defined and constructed determines the degree of interpretative ambiguity. Indicators can be more precise in their meaning when they encapsulate more information or more explicit information in terms of the target inference. Such disambiguation, which counters construct-irrelevant variance, can be achieved by carefully designing the activity space and evidence identification rules. For instance, a time on task indicator does not include much information; it is not self-explanatory, but needs to be interpreted based on additional information such as item and person characteristics (Goldhammer et al., 2014). Another example is the time interval of no interaction as an indicator for the planning state (Eichmann et al., 2019). In contrast, a process indicator capturing the presence of a certain solution strategy encapsulates richer information in the form of a task-specific sequence or pattern of actions and/or states. For instance, in Example 2, the indicator for sourcing during reading has a relatively clear interpretation given the task design and the contextualization included in the evidence identification rules. Of course, claims (e.g., construct interpretations) based on more ambiguous indicators are more ambitious. In such cases, validation becomes more challenging because more extensive support is required in the form of theoretical and empirical evidence.

As described above, continuous assessment (e.g., game-based or simulation-based assessment) and continuous assessment within cognitive items following the standard assessment paradigm (e.g., complex interactive items administered in PISA or PIAAC) exhibit both commonalities and differences. In line with Mislevy (2019), we argued that validation strategies proposed for the standard assessment paradigm are applicable to such complex interactive assessments (e.g., as demonstrated in Example 2). However, validating continuous assessments may be more challenging than continuous assessments within cognitive items, as (evidentiary) validity arguments are likely to be more complex for several reasons. Evidence identification in a game-based or simulation-based assessment is based on a continuous flow of actions and could include more inferential layers due to "evidence-bearing opportunities" (Mislevy, 2019, p. 5) that are not preconstructed in the form of items. The higher complexity of continuous assessments refers not only to the empirical validation phase based on collected data, but also to the design phase, where the construct, behavioral evidence, and activity space must be clearly linked in order to ensure valid inferences. Although continuous or ongoing assessments promise to achieve broad coverage of a construct, ensuring that there is no construct under- or overrepresentation across the encountered activities remains a challenge (DiCerbo et al., 2016), whereas in the standard assessment paradigm, the construct representation is directly determined by the test specification. Relatedly, the construct representation approach, which focuses on differences in item difficulty, is easier to implement when the assessment is structured into items from the start, as in the standard assessment paradigm. As discussed by Mislevy (2019), an evidence-bearing opportunity that is not preconstructed but detected in the course of an ongoing stream of events represents a task situation defined by a state vector showing a certain configuration of values. Thus, this kind of task situation can be understood as a certain row in a Q-matrix mapping the task situation—just as in the case of traditional items—to a set of common characteristics expected to affect the difficulty. Finally, validation strategies that are applied to test-level indicators, such as the nomothetic span approach, can be implemented in a comparable way for continuous assessments and continuous assessments within cognitive items.

Taken together, and as discussed, substantive theories are of great importance for task design, evidence identification, and validation. However, there is a lack of theory or process models relating behavioral low-level features to attributes of the work process through evidence identification and accumulation (Kane & Mislevy, 2017; Mislevy et al., 2012). In particular, such comprehensive theories may not be available for complex ability/competence domains encompassing various cognitive processes. Thus, both exploratory analyses enabling theory development and data-driven approaches informing evidence identification are needed (e.g., Rupp, Levy, et al., 2012). Cumulative evidence is essential for the theory-building process. A necessary condition of this is that the construction of process indicators be transparent (i.e., how derived actions and states are linked to log events is made explicit) and in turn reproducible across studies (Kroehne & Goldhammer, 2018, in press).

Conclusion

Continuous assessment based on log data from complex interactive items makes it possible to derive process indicators representing attributes of the work process. This enables new insights into the work process and underlying cognitive (as well as meta-cognitive, motivational, and affective) constructs. The interpretation of such process indicators needs to be challenged with appropriate validation strategies and should already be considered when designing the task environment, items, and activity space. Thus, the use of log data for assessment purposes should already be considered in the design phase to ensure and improve the valid interpretation of the derived process indicators.

Abbreviations

ECD: Evidence-centered design; IEA: International Association for the Evaluation of Educational Achievement; lltm + e: Linear logistic test model with error term; MDC: Multiple document comprehension; OECD: Organisation for Economic Co-operation and Development; P + : Proportion correct; P + > 0%: Proportion correct greater than 0%; PIAAC : Programme for the International Assessment of Adult Competencies; PIRLS: Progress in International Reading Literacy Study; PISA: Programme for International Student Assessment; RTE: Response time effort; TIMSS: Trends in International Mathematics and Science Study; VI: Visual inspection.

Acknowledgements

Not applicable.

Authors' contributions

FG developed the idea of the article and wrote a first draft. CH, UK and FZ supported the preparation of the manuscript by making suggestions on its structure and line of reasoning as well as by editing parts of the manuscript. All authors read and approved the final manuscript.

Authors' information

A previous version of the article was presented at the conference "Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large Scale Assessments" in Dublin, May 2019, organized by the Educational Research Centre (Dublin, Ireland) and Educational Testing Service (Princeton, NJ, USA).

Funding

This research was supported by the Centre for International Student Assessment (ZIB).

Availability of data and materials

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

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1] DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt/Main, Germany.
[2] Centre for International Student Assessment (ZIB), Frankfurt/Main, Germany.

Received: 26 May 2020 Accepted: 12 September 2021
Published online: 01 October 2021

References

Abele, S., & von Davier, M. (2019). CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models: Models and model extensions, applications, software packages* (pp. 461–488). Springer International Publishing. https://doi.org/10.1007/978-3-030-05584-4_22

AERA, APA, NCME, & Joint Committee on Standards for Educational Psychological Testing. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp. 11–39). Springer. https://doi.org/10.1007/978-3-642-69746-3_2

Aleven, V., Roll, I., Mclaren, B., & Koedinger, K. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educational Psychologist, 45*, 224–233. https://doi.org/10.1080/00461520.2010.517740

Behrens, J. T., & DiCerbo, K. E. (2014). Harnessing the currents of the digital ocean. In J. A. Larusson & B. White (Eds.), *Learning analytics: From research to practice* (pp. 39–60). Springer. https://doi.org/10.1007/978-1-4614-3305-7_3

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP Reading assessment. *Teachers College Record, 113*(11), 2309–2344.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

De Boeck, P., Chen, H., & Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology, 70*(2), 225–237. https://doi.org/10.1111/bmsp.12094

de Klerk, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2015). Psychometric analysis of the performance data of simulation-based assessment: A systematic review and a Bayesian network example. *Computers & Education, 85*, 23–34. https://doi.org/10.1016/j.compedu.2014.12.020

DiCerbo, K. E., Frezzo, D. C., & Deng, T. (2011). Substantive validity of a simulation-based game. *Research and Practice in Technology Enhanced Learning, 6*(3), 161–185.

DiCerbo, K. E., Shute, V., & Kim, Y. (2016). *The future of assessment in technology-rich environments: Psychometric considerations* (pp. 1–21). Springer International.

Drachsler, H., & Goldhammer, F. (2020). Learning Analytics and eassessment—towards computational psychometrics by combining psychometrics with learning analytics. In D. Burgos (Ed.), *Radical solutions and learning analytics: Personalised learning and teaching through big data* (pp. 67–80). Springer Singapore. https://doi.org/10.1007/978-981-15-4526-9_5

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

Eichmann, B., Greiff, S., Naumann, J., Brandhuber, L., & Goldhammer, F. (2020). Exploring behavioural patterns during complex problem solving. *Journal of Computer Assisted Learning, 36*(6), 933–956.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning using examinee response processes for the next generation of assessments*. Routledge.

Gerhand, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory & Cognition, 27*(4), 592–602. https://doi.org/10.3758/BF03211553

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. Doi: https://doi.org/10.1080/15366367.2015.1100020

Goldhammer, F., Hahnel, C., & Kroehne, U. (2020). Analyzing log file data from PIAAC. In D. B. Maehler & B. Rammstedt (Eds.), *Large-scale cognitive assessment: analysing PIAAC data* (pp. 239–269). Springer International Publishing.

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (Vol. 133). OECD Publishing.

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). 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), 18. https://doi.org/10.1186/s40536-017-0051-9

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/10.1037/a0034716

Goldhammer, F., Kroehne, U., Hahnel, C., & De Boeck, P. (2021). 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., & Zehner, F. (2017). What to make of and how to interpret process data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

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

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

Hahnel, C., Kroehne, U., Goldhammer, F., Schoor, C., Mahlow, N., & Artelt, C. (2019). Validating process variables of sourcing in an assessment of multiple document comprehension. *British Journal of Educational Psychology*, *89*(3), 524–537. https://doi.org/10.1111/bjep.12278

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

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., & Mislevy, R. J. (2018). The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design: The evidence trace file. *ETS Research Report Series, 2018*(1), 1–16. https://doi.org/10.1002/ets2.12215

He, Q., & Von Davier, M. (2016). *Analyzing Process Data from Problem-Solving Items with N-Grams: Insights from a Computer-Based Large-Scale Assessment* (pp. 749–776). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch029

Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education, 39*(5), 304–312. https://doi.org/10.1177/0741932518773154

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189–212). Springer.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. T., & Mislevy, R. J. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assesments* (pp. 11–24). Routledge.

Kerr, D., Andrews, J. J., & Mislevy, R. J. (2016). The in-task assessment framework for behavioral data. *The Wiley handbook of cognition and assessment* (pp. 472–507). John Wiley & Sons Ltd. https://doi.org/10.1002/9781118956588.ch20

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

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

Kroehne, U. (2021). *LogFSM: Analysis of log data using finite-state machines.* https://github.com/kroehne/LogFSM

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., & Goldhammer, F. (in press). Tools for analyzing log file data. In L. Khorramdel, M. von Davier, & K. Yamamoto (Eds.), *Innovative computer-based international large-scale assessments—foundations, methodologies and quality assurance procedures*. Springer

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), 8. https://doi.org/10.1186/s40536-014-0008-1

Levy, R. (2020). Implications of considering response process data for greater and lesser psychometrics. *Educational Assessment, 25*(3), 218–235. https://doi.org/10.1080/10627197.2020.1804352

Li, Z., Banerjee, J., & Zumbo, B. D. (2017). Response time data as validity evidence: Has it lived up to its promise and if not, what would it take to do so. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 159–177). Springer International Publishing. https://doi.org/10.1007/978-3-319-56129-5_9

Maddox, B. (2017). Talk and Gesture as Process Data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 113–127. https://doi.org/10.1080/15366367.2017.1392821

Margolis, M. J., & Clauser, B. E. (2006). A regression-based procedure for automated scoring of a complex medical performance assessment. In D. M. Williamson, I. I. Bejar, & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 123–168). Lawrence Erlbaum Associates.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Mislevy, R. J. (2019). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao, R. W. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics* (pp. 1–52). Information Age Publishing.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i–29. https://doi.org/10.1002/j.2333-8504.2003.tb01908.x

Mislevy, R. J., Behrens, J., DiCerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. *Journal of Educational Data Mining, 4*, 11–48.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A. A., Hao, J., Corrigan, S., Hoffman, E., DiCerbo, K. E., & John, M. (2014). *Psychometric considerations in game-based assessment*. GlassLab Research, Institute of Play.

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

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

OECD. (2013). *PISA 2012 assessment and analytical framework: mathematics, reading, science*. OECD Publishing.

OECD. (2017). *PISA 2015 assessment and analytical framework*. OECD Publishing. https://www.oecd-ilibrary.org/content/publication/9789264281820-en

OECD. (2019). *Beyond proficiency: Using log files to understand respondent behaviour in the survey of adult skills*. OECD Publishing.

Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analysing, and interpreting response time, eye tracking and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assesments* (pp. 39–51). Routledge.

Rupp, A. A., Levy, R., Dicerbo, K. E., Sweet, S. J., Crawford, A. V., Caliço, T., Benson, M., Fay, D., Kunze, K. L., Mislevy, R. J., & Behrens, J. T. (2012a). Putting ECD into practice: The interplay of theory and data in evidence models within a digital learning environment. *Journal of Educational Data Mining, 4*(1), 49–110. https://doi.org/10.5281/zenodo.3554643

Rupp, A. A., Nugent, R., & Nelson, B. (2012b). Evidence-centered design for diagnostic Assessment within digital learning environments: integrating modern psychometrics and educational data mining. *Journal of Educational Data Mining, 4*(1), 1–10.

Scalise, K. (2012). Creating innovative assessment items and test forms. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessment: Recent history and predictions for the future* (pp. 133–156). Information Age Publishing.

Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., & Artelt, C. (2020). Multiple document comprehension of university students. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, & C. Lautenbach (Eds.), *Student learning in German higher education: Innovative measurement approaches and research results* (pp. 221–240). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27886-1_11

Shute, V. (2015). Stealth assessment. In J. Spector (Ed.), *The SAGE encyclopedia of educational technology* (pp. 675–676). SAGE Publications Inc.

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

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.

van der Linden, W. J. (2008). Using Response Times for Item Selection in Adaptive Testing. *Journal of Educational and Behavioral Statistics, 33*(1), 5–20. https://doi.org/10.3102/1076998607302626

von Davier, A. A. (2017). Computational psychometrics in support of collaborative educational assessments. *Journal of Educational Measurement, 54*(1), 3–11. https://doi.org/10.1111/jedm.12129

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, *36*(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163–183. https://doi.org/10.1207/s15324818ame1802_2

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*(2), 183–192. https://doi.org/10.1080/08957347.2019.1577248

Yamamoto, K., & Lennon, M. L. (2018). Understanding and detecting data fabrication in large-scale assessments. *Quality Assurance in Education, 26*(2), 196–212. https://doi.org/10.1108/QAE-07-2017-0038

Publisher's Note

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