Original Article

Process and Product in Computer-Based Assessments

Clearing the Ground for a Holistic Validity Framework

Bruno D. Zumbo[1], Bryan Maddox[2,3], and Naomi M. Care[4]

[1]Measurement, Evaluation, & Research Methodology Program, Department of ECPS, University of British Columbia, Vancouver, BC, Canada
[2]School of Education and Lifelong Learning, University of East Anglia, Norwich, UK
[3]Centre for Educational Measurement (CEMO), University of Oslo, Norway
[4]Assessment Microanalytics, The Enterprise Centre, University of East Anglia, Norwich, UK

Abstract: There is no consensus among assessment researchers about many of the central problems of response process data, including what is it and what is it comprised of. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) locate process data within their five sources of validity evidence. However, we rarely see a conceptualization of response processes; rather, the focus is on the techniques and methods of assembling response process indices or statistical models. The method often overrides clear definitions, and, as a field, we may therefore conflate method and methodology – much like we have conflated validity and validation (Zumbo, 2007). In this paper, we aim to clear the conceptual ground to explore the scope of a holistic framework for the validation of process and product. We review prominent conceptualizations of response processes and their sources and explore some fundamental questions: Should we make a theoretical and practical distinction between response processes and response data? To what extent do the uses of process data reflect the principles of deliberate, educational, and psychological measurement? To answer these questions, we consider the case of item response times and the potential for variation associated with disability and neurodiversity.

Keywords: process data, educational assessment, validity, digital assessment

There is much anticipation of the many advantages of computer-based assessments. However, chief among these advantages is the potential of recording and using process data to "open the black box," as the editors of this special issue stated in the call for papers (Lindner & Greiff, 2021), suggesting that one is peering into the brains, thoughts, and minds of test takers as they take a computer-based assessment with the aim to increase the quality of psychological and educational assessment. As one can imagine, this promised journey into the domain of the neuro and cognitive has drawn the attention of many travelers to response processes.

The idea of peering into the black box is closely tied to the earliest descriptions of response processes as validity evidence in the transition from the behaviorist to information processing and early traditions of cognitive psychology seen in the work of Roger Lenon in the mid-1950s, Susan Embretson (Whitely) in the mid-1970s to the present, and Samuel Messick's theoretical developments in the 1980s and 1990s (Hubley & Zumbo, 2017; Messick, 1989, 1995). The term "black box" was meant to describe the opaque mechanics or stages of mental operations that somehow transform inputs into outputs, that is, what goes on between the presentation of the stimulus (the test item or task) as inputs and the test taker's responses as outputs.

Early signs of what has come to be called response processes research included studies by experimental psychologists such as Saul Sternberg (1969), who followed up on a line of research dating back at least 100 years earlier. These experiments focused on elucidating various information processing theories, such as decomposing reaction time distributions for mental tasks such as encoding numerals and their translation into spoken digits that were not unlike those in early intelligence tests or other tests of mental abilities. The aim was to decompose the reaction time into a sequential process between when the item was presented to the respondent (i.e., the stimulus) and their response. Saul Sternberg's (1969, p. 276) report began with a clear statement of the fundamental assumption of information processes approach to response processes dating back at least a hundred years earlier: "[t]he work of Donders (1869) that we have been commemorating was based on the idea that by a train of successive processes, or stages: each component process begins only when the preceding one has ended."

These early signs of information processing research led to a nascent kind of cognitive-psychometric modeling of response processes initiated in the mid-1970s by Susan Embretson (Whitely) (e.g., Embretson, 1983, 1984, 1993; Embretson et al., 1986; Whitely, 1977), James Pellegrino (e.g., Pellegrino et al., 1999, 2016; Pellegrino & Glaser, 1979), Robert Sternberg (1977, 1980), and others who aimed to develop and test formal cognitive-psychometric models of item and test responding in support of test design and validation.

In contrast to the earliest information-processing and cognitive theory building, a current research tradition focuses on process data – with a data-first orientation. In this case, the "response processes" refer to log files, including keystroke data, clickstream data, navigation behaviors, eye-tracking, or video stream, and the associated time stamps, performed by a test-taker to complete a task generated by human-computer interactive items, to name a few (Oranje et al., 2017). When it is attended to, theorizing is an emergent statistical construction that takes second place to the data such as log files. The focus on data and how it is collected can also be seen in the tradition of process research that focuses on interviews and think-aloud methods, which we discuss in more detail later in this paper.

Our concern, however, is that the phrase "response process" has become overused, and it is hard to say exactly when the term broke loose from its information processing moorings or what the reasons were for its subsequent proliferation. Although these reasons are undoubtedly complex and could be the subject of a separate historical study, our concern is prospective, not retrospective – explaining what we mean by response processes.

Response processes are often listed as sources of information to enhance assessment design or validity evidence. However, we rarely see a conceptualization of response processes; rather, the focus is on the techniques and methods of assembling response process indices or statistical models. As such, the method often overrides clear definitions, and, as a field, we may therefore conflate method and methodology – much like we have conflated validity and validation (Zumbo, 2007).

This is not to say that conceptualizations have not been offered in the field. Quite to the contrary, our first aim is to review prominent conceptualizations of response processes to clear the conceptual ground to make it possible for extensive and richer use of response process data to increase test validity. In doing so, we locate process data within the five sources of validity evidence as described in *The Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) and within established theories and practices of validation and consider the validity of its use as evidence of test taker performance. What emerges from these discussions and the use of process data in assessment suggests a set of compelling theoretical and methodological questions about the theoretical and practical distinction between response processes and response data and their implications. Likewise, we address whether there can or should be a dichotomy between process data and product (score) data in terms of how we approach validity.

In the paper, we aim to set out the scope for a holistic framework that can be used as a reference point for analyzing any source of process data. That one recognizes and anticipates multiple sources of process data from a wide array of sensors and measures – some of which may be repeated and automatic, and others (such as cognitive interviews and ethnography) that may be idiosyncratic. In doing so, we aim to provide a more robust theoretical basis for using process data in computer-based assessments.

Our understanding of process data in this paper comes from a hybrid of pragmatism and conceptual analysis that recognizes its multiple sources and evolving uses across the assessment cycle. It also recognizes the transition that is taking place in test design that moves from the treatment of process data as a by-product of computer-based testing to that of process data by design. Maddox (2023) argues that the current uses of process data include: the design and field testing of test items to quality assurance, enhancing the ways that we understand test engagement and performance, and how we validate the interpretation and use of assessment results (Maddox, 2023).

Process and Product

The distinction between process and product in educational and psychological assessment has, until the digital transition, been reasonably stable – product referred to the given answer or solution to an assessment task or item, and a process referred to the way respondents arrived at their answers. When Messick (1995) referred to the use of "direct probes" to investigate response processes, he was talking about how we might gain some insight into the "underlying test responses" that explain how test takers arrive at their answers.

> "At the simplest level, this might involve querying respondents about their solution processes or asking them to think aloud while responding to exercises during field trials" (Messick, 1995, p. 743).

The status and purpose of such probes, models, and evidence of performance were to investigate and support construct validity and the inferences we make from scores derived from test products. It was not a question of viewing data on response processes as a source of test scores.

However, with the digital transition, test designers are increasingly treating data on response processes as a source of information about respondent ability and engagement. That is, to consider data on response processes as sources of evidence to inform or modify test scores. That involves shifting from viewing the product as the primary source of information to inform the test score to the idea that there can be multiple assessment points (i.e., data points) within process-oriented test items. As a result, there is a requirement to consider data on response processes as a source of validity evidence and validate its use as a source of constructing relevant information on aspects of respondent performance (Goldhammer et al., 2021).

There is no widely accepted definition of response processes. Researchers' positionality and methodological preoccupations have largely informed their thinking. Researchers who are invested in collecting and using log data tend to conceptualize response processes in terms of keystrokes and item response times. The qualitative traditions associated with cognitive laboratories have a different tradition of making use of think-aloud protocols and eye-tracking. At the same time, researchers in psychology tend to be more oriented toward the collection and use of physiological data.

That is not to say that researchers have not attempted to define response processes. In their edited volume on the uses of response process data in educational assessment, Ercikan and Pellegrino (2017) suggest the following:

> "Response processes refer to the thought processes, strategies, approaches, and behaviours of examinees when they read, interpret and formulate solutions to assessment tasks." (Ercikan & Pellegrino, 2017, p. 2).

Their concise definition is to be one step removed from process data, and the holistic and pragmatic terms of their definition suggest an apparent conceptual simplicity. However, its different areas of foci concern different phenomena – some of which may be inferred (e.g., thought processes) and others that can more readily be observed and captured (e.g., behaviors). Similarly, they include second-order explanations' strategies, approaches' which might be inferred from a theory of test-taker behavior – what Goldhammer and colleagues (2021) describe as high-level interpretations.

In their edited volume on *Understanding and investigating response processes in validation research*, Zumbo and Hubley (2017) propose a similar definition to that of Ercikan and Pellegrino:

> "... one may think broadly of response processes as the mechanisms that underlie what people do, think or feel when interacting with, and responding to, the item or task and are responsible for generating observed test score variation." (Hubley & Zumbo, 2017, p. 2).

As they note, their definition

> "... expands response process beyond the cognitive realm to include emotions, motivations, and behaviors. Inclusion of affect and motives allows us to take into account how these may impact the different respondents' interactions with the item(s), test, and testing situation. Our definition also requires one to go beyond the surface content of the actions, thoughts, or emotions expressed by, or observed in, respondents to identify the mechanisms that underlie this content. Finally, we encourage researchers and theorists to develop contextualized and dynamic frameworks that take into account the situational, cultural, or ecological aspects of testing when exploring evidence based on response processes." (Hubley & Zumbo, 2017, pp. 2–3).

The Hubley and Zumbo definition locates response processes as *underlying mechanisms* that are somehow revealed by observable data on how people interact with and respond to test items or tasks. Their discussion gives conceptual primacy to clarifying the meaning of response processes rather than process data. Their treatment of response processes seems to position *observed* responses to surface-level features, or what Goldhammer and colleagues (2021) call "low-level features." The distinction between hidden (response processes) and revealed (process data) is also promoted by Ercikan and colleagues (2020) when they argue that:

> "...cognitive responses themselves are not observable. What is captured in the think aloud protocols as well as in log files need to be considered as 'traces of processes' rather than processes themselves." (Ercikan et al., 2020, p. 3).

Therefore, the conceptual distinction between response processes and process data reflects similar thinking in validity theory concerning the difference between observed attributes and underlying traits (Kane, 2009; Kane & Mislevy, 2017). This suggests that process and product could readily be integrated into a holistic validation framework. Indeed, this is what Goldhammer and colleagues (2021) propose as they describe a procedure to validate the interpretation and use of process data, supported by appropriate theory, argument, and evidence:

> "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 et al., 2021, p. 13)

Their argument suggests a requirement to validate the process and product data in the same way under a holistic validity framework that includes a plurality of constructs. So far, so good. However, before integrating process and product, we need to clear the ground by considering questions relating to their different ontological status and characteristics.

Measurement Opportunities

Test item scores based on conventional products reflect the intentionality of design that is informed by principles, theory, evidence, and an established architecture of procedures in psychological and educational measurement. In contrast, the use of process data as a "by-product" of computer-based testing is necessarily opportunistic, viewing technological interfaces and sensors as convenient measurement opportunities. In contrast, we may consider the use of process data "by design" as closer to the deliberate rationales of products. As we argue, response processes involve not only the cognitive strategies and approaches of test takers but also emotion, affect, interaction, physiology, and embodied behavior in the test ecology.

As a result, there is no reason why the diverse sources of process data – from physiological responses to patterns of gaze and keystrokes should necessarily provide valid evidence of variation in the test construct. Before we adopt a holistic validity framework that incorporates product and process in assessment, we must therefore recognize that those sources of data are not equivalent. That is, sources of process data should not be considered as a de-facto source of construct relevant evidence, in the way that Embretson (2016) suggests in her discussion of "nomothetic span" or the wider sources of validation evidence as argued by Cronbach and Meehl's (1955) in their discussion of a "nomological net."

The status of process data as one of the five "sources" of validity evidence in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) may have neglected the requirement to validate the uses of process data. The oft-assumed truthfulness and precision of probes and measures associated with digitization have added to that sense of apparent objectivity in using process data to investigate validity (see Aryadoust et al., 2022; Oranje et al. 2017; Yaneva et al., 2021).

As the uses of process data expand across the assessment cycle, including high-stakes inferences about test taker engagement and performance, quality assurance and interpretations of video data for remote proctoring, there is a need to explicitly and systematically establish the validity of its use and its reliability and fairness for use across diverse contexts and populations. Nevertheless, most of the discussions of process data are currently framed in terms of exploratory methodological studies. This lack of appropriate validation is evident even in some of the most established techniques as think-aloud protocols, whose use goes back to the 1950s (Leighton et al., 2017; Padilla & Leighton, 2017).

That is not to say that sources of process data should not be used for assessment purposes. On the contrary, their potential to inform and improve digital assessments is clear (Goldhammer et al., 2020; Jiao et al., 2021). However, in the absence of systematic and routinized validation, there is a requirement for more explicit arguments and evidence than we might consider necessary for product-based scores (Goldhammer et al., 2021).

The Plurality of Constructs and Validation Methods

As Shear and Zumbo (2014) show, over the last nearly 60 years, the concepts and theories of test validity have grown increasingly expansive, and the methods for test validation have become increasingly complex and multi-faceted. Validity theorists have highlighted the important distinction between validity and validation (Zumbo, 2007). Validity is the property or relationship we are trying to judge; validation is an activity geared toward understanding and making that judgment. Zumbo (2009) reminds us of the importance that a guiding rationale (i.e., validity theory) must play in selecting and applying appropriate validation research methods (i.e., validation) and points up the importance of having a clear concept of validity.

Following the positionality of "construct theory" embraced by Goldhammer and colleagues (2021), this need for clarity of the concept of validity is paramount within a holistic framework where one can conceive of the test-oriented construct(s) and process-oriented constructs. It should be noted that many authors refer to construct validity as the most important characteristic of a test, but it is seldom defined. A clear statement of what a construct is and the logic of construct validation was presented by Cronbach and Meehl (1955). These authors wrote:

> "A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute… or to possess some degree of a quantitative attribute… Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized

> or loose, used in ramified theory or in a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test…" (p. 247)

Informed by Cronbach and Meehl's (1955) description, we consider that the attribute(s) of the test taker about which we make claims in interpreting a test include:

- "Test-construct(s)," reflected in the product-oriented test-taker response data to the items or tasks in the test (i.e., their deliberate and conscious responses), and;
- "Process-construct(s)," reflected in the process data recorded by the various sensors and stored in the log files, including keystroke data, eye-tracking, and the associated time stamps, performed by a test-taker as they complete a task.

In the encounter between the test taker and the test, we may make inferences about the test construct via formal responses to an item (or task), as well as process-oriented data that are intended to be used to replace or supplement the conventional product. That approach is evident, for example, in process-oriented assessments of problem-solving and assessments of clinical decision-making. The item responses are then scored and validated in a way consistent with the intended test construct to be measured. In contrast, process data may also be used to support important inferences about some other process-related construct, such as engagement, effort or motivation, user experience, or item accessibility.

Importantly, a holistic validity framework recognizes that digital process data, despite its apparent accuracy and interpretation, is fallible, whether used to support claims made about test constructs or process-oriented constructs (Maddox, 2017). From the perspective of a holistic validity framework, measures of process data may therefore contribute to sources of construct irrelevant variance, construct underrepresentation, or conditions that create a lack of measurement invariance that confounds the claims made about the status of the test takers regarding the construct or attribute of interest.

Validating Claims Made From Test-Constructs and Process-Constructs

There are currently few professionally established procedures and tests to validate the use of process data in assessment. When Ercikan and colleagues (2020) describe "differential response time" and "differential response processes," they invoke established concepts and methods in the product-oriented analysis of differential item functioning (DIF):

> "For identifying differential response time (DRT) for the ELL and non-ELL groups, we used an extension of Differential Item Functioning (DIF). DIF analysis is a standard fairness practice in the testing industry since the 1980s." (2020, p. 7)

The logic of such an application of DIF analysis is clear, and this would certainly support a move toward a holistic validity framework for process and product in assessment. However, it remains unclear whether data on response times behave in ways equivalent to product-oriented scores and to what extent they might require different treatment. Indeed, in the conclusion of their paper, Ercikan and colleagues (2020) acknowledge the need for further work in this area to strengthen their inferences with a broader set of information about how respondents navigate test items. Of the potential threats to the validity of process data, the most commonly acknowledged is its partiality. As a result, most accounts of process data and its validation acknowledge the benefit of triangulation with multiple sources of process and product data (Ercikan et al., 2020; Goldhammer et al., 2021; Li et al., 2017; Maddox, 2017; Maddox & Zumbo, 2017; Maddox et al., 2019; Oranje et al., 2017). Log data, in particular, are recognized as a partial account of response processes. While computers can capture each interaction with the keyboard and mouse as "events" with accuracy to the millisecond, they are unable to account for "idle time" in the data when the respondent looks at the screen, or interacts with the test administrator, or uses a scratch pad or off-screen calculator (Salles et al., 2020).

As we have argued above, there is no particular reason why all sources of process data would necessarily exhibit construct-relevant variation. Conventional DIF analysis examines the source of construct irrelevant variation that may be associated with responses to test item content and design, personal characteristics, or some wider aspect of the testing situation or its wider ecology (Chen & Zumbo, 2017; Zumbo, 2007; Zumbo et al., 2015). It goes without saying that those sources of construct irrelevant variation may also be present in process data. However, a further challenge with process data is to investigate potential measurement bias (and potential unfairness) of the data arising from sensors. This challenge is especially pressing when the sources of process data such as keystrokes, gaze, or physiological measures (e.g., electrodermal activity, pupil diameter, facial expressions, neuroimaging, and measures of heart rate) are re-purposed for assessment purposes.

Item Response Times and Disability

The use of response time data is one of the most established applications of response process data in assessment, with a degree of formalization about its methods (Lee & Jia, 2014; Li et al., 2017; Wise, 2019) and related discussion of validation (Goldhammer et al., 2021; Li et al., 2017; Reis Costa et al., 2021; Wise, 2017, 2019). The digital transition has witnessed extensive application, such as measures to support inferences about rapid guessing behaviors and disengagement in large-scale assessments (Ercikan et al., 2020; Goldhammer et al., 2017; Kroehne et al., 2020; Lundgren & Eklöf, 2020; Soland et al., 2018; Wise, 2017, 2019, 2020a, 2020b; Wise & Gao, 2017, Wise & Kong, 2005, Wise et al., 2019), and to explore the relationship between response times, accuracy and ability (e.g., Ivanova et al, 2020; Michaelides et al., 2020; Ranger et al., 2021; Reis Costa et al., 2021).

However, despite their apparent accuracy, response times alone offer little information about the actual response processes of test takers, in the sense of their thought processes, strategies, and behaviors (Li et al., 2017; Maddox et al., 2019). As a result, the uses of response time data often lack wider justification for its interpretive claims about test-taking behavior – such as claims about the detection of "rapid disengaged guessing." There is little discussion, for example, of the presence of "slow" guessing or threats to measurement precision in the variation of item load times. Furthermore, the literature on the use of response time data in assessment, with its built-in assumptions of measurement precision, rarely takes time to establish invariance of the claims made from data arising from sensors.

Two points are noteworthy. First, in what is still considered by many as the seminal treatise on response time in cognitive and experimental research (Luce, 1986, pp. 173–174), the eminent mathematical psychologist Duncan Luce was *not as sanguine* about deriving underlying mental processes from the distribution of response time alone by mathematically or statistically modeling to decompose response time distributions that reflect different cognitive processes. In particular, he expressed concern that the relationship between mental processes and the resultant response time data distribution is not as clear or discernible as one would like to allow researchers to easily differentiate between the component response time distributions that reflect the different response processes. Second, concerns similar to those stated by Luce about the decomposition of distributions of response times alone into successive cognitive processes or stages would also apply to other sensor-based information on response processes, including the interpretation of eye movements or physiological measures of affective orientation and "somatic markers" (Damasio, 2018; Asma & Gabriel, 2019) such as variations in a heartbeat, data from electrodermal activity or pupil dilation. Technical advances, with the capture of log files (the interface with the keyboard and mouse), and other bodily sensors enable the capture of data on response processes (Aryadoust et al., 2022). However, we cannot assume that such measures, which concern embodied chemical and biological processes (e.g., the role of endorphins, adrenalin etc.) are of the same type as keystrokes, or indeed, that they are unrelated. To take "embodied" seriously means to consider their neurological and chemical basis, as well as the social and ecological significance of context and the "extended mind" (Clark, 2011), whether it involves virtual phenomena in onscreen interactions or the wider significance of the testing situation (e.g., setting, time, stakes).

To further explore these themes, we will consider the significance of disability and neurodiversity in tested populations. There is a broad diversity within human neurobiology (Pellicano & den Houting, 2022); the human brain develops and functions in countless ways, resulting in a test-taking population with diverse strategies and responses. There is a need to recognize that, rather than anomalies, test-takers with disabilities and learning differences represent a sizeable minority. Using the UK Equality Act, recent prevalence data estimates that 21% of the UK population has a disability (Kirk-Wade, 2022). Such estimates are unlikely to include many test-takers with forms of neurodiversity or learning differences that do not meet the threshold for disability. Disabilities and neurodiversity can lead to test takers responding to test items in ways that deviate from established models.

It has long been recognized that students with disabilities can be disadvantaged by poor item design or test quality that can unfairly lead to DIF (Camilli, 2013; Sireci, 2014). For example, a recent study conducting a DIF analysis on the Central Examination for Secondary Education Institutions in Turkey revealed that 5.56% of the total items created an unfair advantage for students without visual impairments compared to students with visual impairments (Şenel, 2021). Even in this product-oriented discussion, the author notes that studies investigating DIF against test-takers with disabilities often split the test-takers into "disabled" and "non-disabled" groups, without taking into account the varying nature of the disabilities or how these may affect the individual (Şenel, 2021). Thus, there is often an implicit assumption that test-takers with disabilities will behave and respond to test items in a unitary fashion.

Since the characteristics and response processes of test-takers with disabilities can be sources of DIF, it is not a considerable leap to accept that test-takers with disabilities may respond to test items in ways that differ from how the typical majority would be reflected in item response times. Furthermore, each different form of neurodiversity, learning difficulty or disability may result in differing responses to item content. This is particularly pertinent when considering the valid interpretation of log data on response times, which often relies on unitary assumptions regarding the cognitive processes and motivation of the test taker (Goldhammer et al., 2021). Such assumptions are tested by disability literature that demonstrates a wide array of cognitive processes and attentional differences that affect approaches to educational tasks. For example, eye-tracking research has shown that individuals with dyslexia tend to have poorer comprehension than controls when reading written information and consistently demonstrate longer reading times, particularly when sentences are ambiguous (Stella & Engelhardt, 2019). It is also well established that individuals with attention deficit/(hyperactivity) disorder (ADHD) have difficulty concentrating and are frequently distracted by task-irrelevant stimuli. In contrast, the opposite pattern is seen in individuals with autism spectrum disorder (ASD), who tend to fixate on a single task (Landry, 2021). Each cognitive variation will, in turn, lead to variations in process data and log times that may, in turn, undermine standard interpretations regarding its validity.

To demonstrate this point, we can consider the response times of individuals with ADHD. In both behavioral and neurocognitive tasks, individuals with ADHD have demonstrated they are "consistently inconsistent" (Kofler et al., 2013). Intra-individual variability (IIV) is considered characteristic of ADHD (Bluschke et al., 2021). That describes fluctuations in task performance and occurs independently of mastery of a given skill (Bluschke et al., 2021). These fluctuations in task performance occur over periods of seconds or milliseconds rather than hours or throughout the day (Kofler et al., 2013). Intra-individual variability can affect many test-taking activities, including handwriting (Borella et al., 2011) and sustained attention (de Zeeuw et al., 2008). When tested in laboratory settings, individuals with ADHD react too early *or* too late to stimuli on computerized tasks, leading to fewer correct responses (Bluschke et al., 2021). Other studies have noted the unpredictability of responses given by individuals with ADHD compared to controls (Aase & Sagvolden, 2005). As such, variance, unpredictability, and "inconsistency" are considered hallmarks of ADHD. These sources of naturally occurring variation in item responses demonstrate the need for the validation of timestamp data on response times to more thoroughly consider sources of variation in response processes beyond guessing and disengagement and to more rigorously explore the limitations of (implicit) assumptions of invariant claims from data arising from sensors. Rather than viewing such variation as a threat to process data validation, it may be viewed as an opportunity to understand further the learning strategies of a range of test takers and improve inclusive test design.

Conclusions

In conclusion, we aim to show the similarity and essential differences between response process theories and then turn to some closing remarks.

Comparing and Contrasting Theories of Response Processes

Figure 1 depicts the space between a test question or task presented to the test-taker and when they respond, highlighting response processes and process data. In behaviorist language that shaped early assessment and testing theories, this question or task is described as the "stimulus (S)." The response to the item or task is the response (R) in that stimulus-response (S-R) view of behavior. Response processes happen in the space between S and R (with task decomposition that can be envisaged and studied as a series of recurrent behaviors, steps, or stages). How we address the question (and empirical problem) of what goes on between stimulus and response is what allows us to compare and contrast various views of response processes throughout the recent history of testing and to describe our holistic framework.

Cronbach and Meehl (1955) acknowledge this S-R space by invoking earlier concepts of intervening variables and hypothetical constructs (MacCorquodale & Meehl, 1948). The later information processing and cognitive psychologist conceived this space as holding the mental processes. Messick referred to mental probes (e.g., think aloud methods) to access this space. In line with the cognitive and information processes tradition, Ercikan and Pelligrino refer to this space as containing thought processes, strategies, approaches, and behaviors of examinees when they read, interpret and formulate solutions to assessment tasks. Hubley and Zumbo take a broader view and expand response processes as mechanisms that underlie what people do, think or feel when interacting with items and responding to them. Therefore, the response processes space is extended beyond the cognitive realm to include emotions, motivations, and behaviors as the mechanisms that underlie and are responsible for generating observed item and test score variation. Zumbo and colleagues (2015) introduce an ecological model of item responding that incorporates a more expansive theory that includes contextual influences reflecting the test takers lived experience and family setting, as well as larger community (and potentially) national characteristics that may influence item responding. In a related vein, adopting some of the language and concepts of generalizability theory, Araneda and colleagues (2022, pp. 2–3) state that "[r]esponse process data can take a variety of forms, only some of which

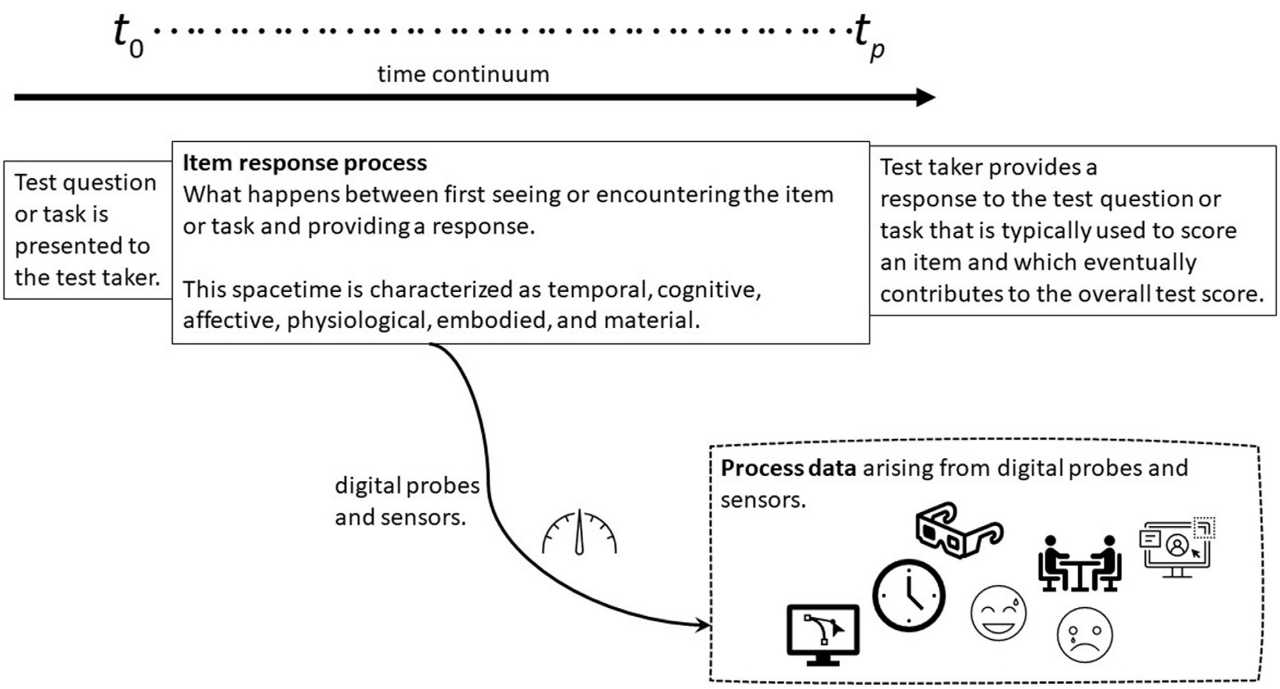

Figure 1. A depiction of the space between when the test question or task is presented to the test-taker and when they respond, highlighting response processes and process data.

are relevant to the constructs intended to be measured . . . there are many ways in which. These processes could be categorized . . . only some are considered to reflect the targeted construct, and typically only a subset of these are captured in scoring." Finally, building upon developments by Maddox, as depicted in Figure 1, we characterize this space as temporal, cognitive, affective, physiological, embodied, and material features.

The essential differences between the theories and viewpoints described above reflect the breadth and scope of characterizations of response processes and the terrain of future research. Some early views conflated what response processes are with how they are attained, for example, Messick characterizes response processes arising from mental probes. Other theories conceive of response processes as mostly cognitive and physiological, wherein the intervening variables are the unobserved mechanics of the "process" leading to the response. Our proposed holistic framework, therefore, articulates a definition and relation between test constructs and process constructs, highlighting the need to rigorously conceptualize and validate the way that response processes and "process data" are treated as measurement opportunities

Closing Remarks

Newton asked if we were on the "brink of a revolution" involving the use of process data in validation research and to what extent it might become part of the routine practice of validation research (2019, p. 246). The subsequent rapid growth of literature on the uses of process data in assessment demonstrates that this indeed was the case (Goldhammer et al., 2020; Jiao et al., 2021). Clearly, data on assessment response processes from digital probes and sensors (including log files) are now central to digital assessment design, even if the field is not yet entirely sure about the kinds of measurement opportunities they involve. The many uses of process data in assessment have become routine, not only for validation purposes but also for wider applications across the assessment cycle. For example, process data such as item response time can be used as ancillary statistical information in calculating person parameter (so-called theta) values reported from an item response theory (IRT) analysis. Alternatively, one can envision an assessment where the process construct is the primary interest, for example, in tests of collaborative problem-solving or clinical decision-making.

In this paper, we attempt to move the argument from one about the use of process data for validation purposes to recognize the wider uses of process data across the assessment cycle. In that sense, we suggest that the conceptualization of process data in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) should be updated to capture the diversity of sources and uses of process data and to acknowledge the associated complexity of explanation.

Whereas the validation of conventional product data is supported by routine protocols and procedures – including public-oriented arguments (Addey et al., 2020) about its fairness and reliability, the uses of process data are yet to establish such equivalence. We, therefore, support the argument of Goldhammer and colleagues (2021) about the need to establish appropriate validity theory, evidence, and arguments about the process and product data.

In clearing the ground for a holistic validity framework, we have highlighted the need to validate the interpretation and use of process data to make inferences about test-oriented and process-oriented constructs and recognize the different ontological statuses of process and product data in assessment. As we have argued, it is misleading to interpret the characteristics of process data – such as the precision of digital measurement in log files and response times, as de-facto evidence of its validity. Indeed, just as unmeasured nuisance variables may influence product data, process data can contain the signatures of construct irrelevant variance, differential sensor functioning or construct under-representation. This is illustrated in our example of disability and neurodiversity. We have shown that the valid interpretation and use of timestamp data – for test-oriented or process-oriented constructs requires appropriate consideration of diversity in what it means to be human.

References

Aase, H., & Sagvolden, T. (2005). Moment-to-moment dynamics of ADHD behaviour. *Behavioral and Brain Functions, 1*(1), 1–14. https://doi.org/10.1186/1744-9081-1-12

Addey, C., Maddox, B., & Zumbo, B. D. (2020). Assembled validity: Rethinking Kane's argument-based approach in the context of International Large-Scale Assessments (ILSAs). *Assessment in Education: Principles, Policy & Practice, 27*(6), 588–606. https://doi.org/10.1080/0969594X.2020.1843136

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Araneda, S., Lee, D., Lewis, J., Sireci, S. G., Moon, J. A., Lehman, B., Arslan, B., & Keehner, M. (2022). Exploring relationships among test takers' behaviors and performance using response process data. *Education Sciences, 12*(2), 104. https://doi.org/10.3390/educsci12020104

Aryadoust, V., Foo, S., & Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Language Testing, 39*(1), 56–89. https://doi.org/10.1177/02655322211026876

Asma, S., & Gabriel, R. (2019). *The emotional mind*. Harvard University Press.

Bluschke, A., Zink, N., Mückschel, M., Roessner, V., & Beste, C. (2021). A novel approach to intra-individual performance variability in ADHD. *European Child & Adolescent Psychiatry, 30*(5), 733–745. https://doi.org/10.1007/s00787-020-01555-y

Borella, E., Chicherio, C., Re, A. M., Sensini, V., & Cornoldi, C. (2011). Increased intraindividual variability is a marker of ADHD but also of dyslexia: A study on handwriting. *Brain and Cognition, 77*(1), 33–39. https://doi.org/10.1016/j.bandc.2011.06.005

Camilli, G. (2013). Ongoing issues in test fairness. *Educational Research and Evaluation, 19*(2–3), 104–120. https://doi.org/10.1080/13803611.2013.767602

Chen, M. Y., & Zumbo, B. D. (2017). Ecological framework of item responding as validity evidence: An application of multilevel DIF modeling using PISA data. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 53–68). Springer.

Clark, A. (2011). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Cronbach, L., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Damasio, A. (2018). *The strange order of things: Life, feeling and the making of cultures*. Pantheon Press.

de Zeeuw, P., Aarnoudse-Moens, C., Bijlhout, J., König, C., Post Uiterweer, A., Papanikolau, A., Hoogenraad, C., Imandt, L., de Been, D., Sergeant, J. A., & Oosterlaan, J. (2008). Inhibitory performance, response speed, intraindividual variability, and response accuracy in ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry, 47*(7), 808–816. https://doi.org/10.1097/CHI.0b013e318172eee9

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

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186. https://doi.org/10.1007/BF02294171

Embretson, S. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Erlbaum.

Embretson, S. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice, 35*(3), 6–22. https://doi.org/10.1111/emip.12117

Embretson, S., Schneider, L. M., & Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23*, 13–32. https://doi.org/10.1111/j.1745-3984.1986.tb00231.x

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

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational Assessment, 25*(3), 179–197. https://doi.org/10.1080/10627197.2020.1804353

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

Goldhammer, F., Scherer, R., & Greiff, S. (2020). Editorial: Advancements in technology-based assessment: Emerging item formats, test designs, and data sources. *Frontiers in Psychology, 10*, Article 3047. https://doi.org/10.3389/fpsyg.2019.03047

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), 1–25. https://doi.org/10.1186/s40536-017-0051-9

Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-319-56129-5_1

Ivanova, M., Michaelides, M., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation, 26*(5–6), 252–274. https://doi.org/10.1080/13803611.2021.1963939

Jiao, H., He, Q., & Veldkamp, B. P. (2021). Editorial: Process data in educational and psychological measurement. *Frontiers in Psychology, 12*, Article 793399. https://doi.org/10.3389/fpsyg.2021.793399

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 39–64). IAP Information Age Publishing.

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

Kirk-Wade, E. (2022). *UK disability statistics: Prevalence and life experiences*. House of Commons Library, Commons Library Research Briefing.

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review, 33*(6), 795–811. https://doi.org/10.1016/j.cpr.2013.06.001

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling, 62*(2), 147–177.

Landry, P. (2021). A behavioral economic theory of cue-induced attention- and task-switching with implications for neurodiversity. *Journal of Economic Psychology, 86*, Article 102423. https://doi.org/10.1016/j.joep.2021.102423

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

Leighton, J. P., Tang, W., & Guo, Q. (2017). Response processes and validity evidence: Controlling for emotions in think aloud interviews. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 137–158). Springer. https://doi.org/10.1007/978-3-319-56129-5_8

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. 159–178). Springer. https://doi.org/10.1007/978-3-319-56129-5_9

Lindner, M. A., & Greiff, S. (2021). Call for Papers "Process data in computer-based assessment: Opening the Black Box" A special issue of the *European Journal of Psychological Assessment*. *European Journal of Psychological Assessment, 37*(2), 161–162. https://doi.org/10.1027/1015-5759/a000658

Luce, R. D. (1986). *Response times*. Oxford University Press.

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*(5–6), 275–301. https://doi.org/10.1080/13803611.2021.1963940

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55*(2), 95–107. https://doi.org/10.1037/h0056029

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

Maddox, B. (2023). *The uses of process data in large-scale educational assessments* (Working Paper). OECD.

Maddox, B., Keslair, F., & Javrh, P. (2019). Investigating testing situations. In B. Maddox (Ed.), *International large-scale assessments in education* (pp. 157–172). Bloomsbury.

Maddox, B., & Zumbo, B. D. (2017). Observing testing situations: Validation as jazz. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 179–192). Springer. https://doi.org/10.1007/978-3-319-56129-5_10

Messick, S. (1989). Validity. In R. K. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan Publishing.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses as scientific enquiry into score meaning. *American Psychologist, 50*, 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Michaelides, M., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187–205. https://doi.org/10.1080/15305058.2019.1706529

Newton, P. E. (2019). What is response process validation evidence and how important is it? An essay reviewing Ercikan and Pellegrino (2017) and Zumbo and Hubley (2017). *Assessment in Education: Principles, Policy & Practice, 26*(2), 245–253. https://doi.org/10.1080/0969594X.2018.1447909

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

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211–228). Springer. https://doi.org/10.1007/978-3-319-56129-5_12

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24*(1), 307–353. https://doi.org/10.3102/0091732X024001307

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550

Pellegrino, J. W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. *Intelligence, 3*(3), 187–215. https://doi.org/10.1016/0160-2896(79)90017-5

Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from "normal science" to neurodiversity in autism science. *Journal of Child Psychology and Psychiatry, 63*, 381–396. https://doi.org/10.1111/jcpp.13534

Ranger, J., Kuhn, J.-T., & Pohl, S. (2021). Effects of motivation on the accuracy and speed of responding in tests: The speed-accuracy tradeoff revisited. *Measurement: Interdisciplinary Research and Perspectives, 19*(1), 15–38. https://doi.org/10.1080/15366367.2020.1750934

Reis Costa, D., Bolsinova, M., Tijmstra, J., & Andersson, B. (2021). Improving the precision of ability estimates using time-on-task variables: Insights from the PISA 2012 computer-based assessment of mathematics. *Frontiers in Psychology, 12*, Article 579128. https://doi.org/10.3389/fpsyg.2021.579128

Salles, S., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: Process data analysis in large-scale mathematics assessment in France. *Large Scale Assessments in Education, 8*, Article 7. https://doi.org/10.1186/s40536-020-00085-y

Şenel, S. (2021). Assessing measurement invariance of Turkish "Central Examination for Secondary Education Institutions" for visually impaired students. *Educational Assessment, Evaluation & Accountability, 33*(4), 621–648. https://doi.org/10.1007/s11092-020-09345-5

Shear, B. R., & Zumbo, B. D. (2014). What counts as evidence: A review of validity studies in educational and psychological measurement. In B. D. Zumbo & E. K. H. Chan (Eds.), *Validity and validation in social, behavioral, and health sciences* (pp. 91–111). Springer. https://doi.org/10.1007/978-3-319-07794-9_6

Sireci, S. G. (2014). No more excuses: New research on assessing students with disabilities. *Journal of Applied Testing Technology, 10*(2), 1–18. https://jattjournal.net/index.php/atp/article/view/48352

Soland, J., Wise, S. L., & Gao, L. (2018). Identifying disengaged survey responses: New evidence using response time metadata. *Applied Measurement in Education, 32*(2), 151–165. https://doi.org/10.1080/08957347.2019.1577244

Stella, M., & Engelhardt, P. E. (2019). Syntactic ambiguity resolution in dyslexia: An examination of cognitive factors underlying eye movement differences and comprehension failures. *Dyslexia, 25*(2), 115–141. https://doi.org/10.1002/dys.1613

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84*, 353–378. https://doi.org/10.1037/0033-295X.84.4.353

Sternberg, R. J. (1980). Representation and process in linear syllogistic reasoning. *Journal of Experimental Psychology, 109*, 119–159. https://doi.org/10.1037/0096-3445.109.2.119

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, 30*, 276–315.

Whitely (Embretson), S. E. (1977). Information-processing on intelligence test items: Some response components. *Applied Psychological Measurement, 1*, 465–476. https://doi.org/10.1177/014662167700100402

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. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325–336. https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L. (2020a). The impact of test-taking disengagement on item content representation. *Applied Measurement in Education, 33*(2), 83–94. https://doi.org/10.1080/08957347.2020.1732386

Wise, S. L. (2020b). Six insights regarding test-taking disengagement. *Educational Research and Evaluation, 26*(5–6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

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

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer based tests. *Applied Measurement in Education, 30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

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

Yaneva, V., Clauser, B. E., Morales, A., & Paniagua, M. (2021). Using eye-tracking data as part of the validity argument for multiple-choice questions: A demonstration. *Journal of Educational Measurement, 58*(4), 515–537. https://doi.org/10.1111/jedm.12304

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, *Vol. 26: Psychometrics* (pp. 45–79). Elsevier.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). IAP Information Age Publishing.

Zumbo, B. D. & Hubley, A. M. (Eds.). (2017). *Understanding and investigating response processes in validation research*. Springer. https://doi.org/10.1007/978-3-319-56129-5

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136–151. https://doi.org/10.1080/15434303.2014.972559

History

Received February 13, 2022
Revision received July 25, 2022
Accepted October 11, 2022
Published online March 21, 2023
EJPA Section / Category Methodological Topics in Assessment

Open Science

Open Data: Due to this paper's conceptual and philosophical nature, there are no data and statistical analysis typically reported in empirical psychological studies.

Open Materials: Due to this paper's conceptual and philosophical nature, there are no research methods and materials typically reported in empirical psychological studies.

Preregistration of Studies and Analysis Plan: Due to this paper's conceptual and philosophical nature, there was no method and analysis plan typically used in empirical psychological studies to preregister.

Funding

This work has been supported by the Canada Research Chairs (AWD-016179 UBCEDUCA 2020) and the UBC Paragon Research Initiative to Bruno Zumbo.

ORCID

Bruno D. Zumbo
https://orcid.org/0000-0003-2885-5724

Bruno D. Zumbo

Department of ECPS
Faculty of Education
University of British Columbia
2125 Main Mall
Vancouver, BC V6T 1Z4
Canada
bruno.zumbo@ubc.ca