

Re-interpreting aberrant responses in educational and psychological assessments

Bryan Maddox & Bruno D. Zumbo

To cite this article: Bryan Maddox & Bruno D. Zumbo (25 Jun 2026): Re-interpreting aberrant responses in educational and psychological assessments, Assessment in Education: Principles, Policy & Practice, DOI: [10.1080/0969594X.2026.2682489](https://doi.org/10.1080/0969594X.2026.2682489)

To link to this article: <https://doi.org/10.1080/0969594X.2026.2682489>



Published online: 25 Jun 2026.



Submit your article to this journal [↗](#)



Article views: 81



View related articles [↗](#)



View Crossmark data [↗](#)



Re-interpreting aberrant responses in educational and psychological assessments

Bryan Maddox ^a and Bruno D. Zumbo ^b

^aDigital Education Futures Initiative, University of Cambridge, Cambridge, UK; ^bCanada Research Chair in Psychometrics and Measurement, Faculty of Education, University of British Columbia, Vancouver, Canada

ABSTRACT

In this paper, we review the notion of aberrant responses in educational and psychological assessments. We examine the literature on aberrant responding, from the early work of Lee Cronbach, to contemporary interpretations of aberrant response processes in digital assessments. We contrast approaches that portray aberrant responses as a de-facto threat to test integrity and validity (i.e. something to be detected and removed), with an alternative paradigm, informed by the work of Cronbach, which interprets anomalous and unexpected responses as a source of formative insight, with potential to improve understanding of test content and design, to illuminate the diverse ways in which tested populations engage with tests and questionnaires. We argue that the latter approach provides stronger foundations for psychometrics and for AI-based models of item responding.

ARTICLE HISTORY

Received 21 April 2025
Accepted 25 May 2026

KEYWORDS

Cronbach; aberrant;
response processes;
response models

Introduction

This paper proposes a new paradigm for the treatment of aberrant responses and aberrant responding in educational assessment. The contemporary significance of the paper lies in the expanded use of digital process data in assessment and its application in AI-based adaptive systems. It involves a reappraisal of the status of aberrant responses in areas such as the use of process-oriented test constructs, test validation and quality assurance, and remote proctoring (Goldhammer et al., 2021; Zumbo et al., 2023). In that context, the ways we interpret and explain the presence of anomalous test responses and response processes have increased importance, which is amplified by their integration into AI-based models (Belzak et al., 2025; Bulut, Beiting-Parrish, et al., 2024; Bulut, Gorgun, et al., 2024; Yan et al., 2024).

The paper is informed by a programme of research that has sought to investigate and re-interpret aberrant (anomalous) responses in order to prompt and support the development of a new generation of theories and methods that can better capture the insights

that they may provide into the factors that inform test performance, test validity, integrity and inclusion.

In this paper, we review the research literature on aberrant responding, from Cronbach's (1946, 1957) early work on 'response sets' to contemporary discussions of aberrant response processes in digital data. By considering different approaches to anomalous responses and response processes, we present an argument and an alternative research paradigm that treat aberrant responses as a source of formative insight rather than as a *de facto* threat to test integrity. We consider that perspective to be more future-ready as we develop new uses of artificial intelligence for immersive, interactive, and personalised assessments.

We do not deny that data on response processes can be used to identify phenomena such as disengagement, data fabrication, pre-knowledge, and cheating. However, we suggest that a high burden of proof is required to avoid measurement error and the negative consequences of wrongly classifying legitimately diverse responses and respondents. The implication is that atypical responses should not be automatically 'flagged' as evidence of test-taker deviance, but rather, to signal the requirement for deeper enquiries into sources of variation – in keeping with the explanation-oriented principles of 3rd Generation DIF and explanation-focused validity (Zumbo, 2007, 2023).

Our suggested response is to parse investigations of response processes into two steps. The first step is to identify atypical or unexpected responses. That step is intended to be agnostic and impartial, avoiding attempts to explain atypical behaviour or locate its cause. A second step, involving an expanded space of enquiry, would involve developing diversity-oriented, multi-parameter theories and methods to investigate, explain and apply the insights generated by anomalous and unexpected responses.

The structure of the paper is as follows. We begin with 'Aberrant Responding' by reviewing and critiquing the different ways the assessment literature constructs and responds to aberrant responses. Those have often portrayed aberrant responses as some negative characteristic of test takers. We contrast that explanation with the proposal of Cronbach (1946, 1957), whose work on 'response sets' points in another direction at identifying some unintended interaction between test takers and test item design, content or administration. In the second part, on 'Formative Insight', we set out a contrasting approach to the analysis of anomalous responses. We propose guiding ideas and principles that inform our argument, including a commitment to use process models that anticipate and embrace diversity as a feature of human populations. In particular, we discuss the implications of our approach in relation to the contemporary literature on response processes and its treatment of aberrant responses among neuro-diverse and disabled test takers.

Aberrant responding has long occupied an uneasy place in the assessment literature, frequently treated as a nuisance to be detected, controlled, or removed in the service of preserving score accuracy. As we show in this paper, many researchers locate the source of aberrant responses within the test taker, attributing them to carelessness, disengagement, low ability, or intentional distortion. In contrast, the perspective that we advance positions aberrant responses as potentially informative signals – clues to underlying interactions among item characteristics, testing conditions, and the diverse ways individuals interpret and engage with assessment tasks. By shifting the focus from test-taker deficiency to the

broader ecology of the testing situation, we reconsider what aberrant responding represents and how it might be leveraged to deepen our understanding of test functioning.

In this paper, we use the term ‘aberrant’ to refer to responses and response processes in assessments and surveys that are thought to contradict the characteristics and norms of response models or the researcher’s assumptions about expected and legitimate responses or response processes. We treat the terms ‘aberrant’ and ‘deviant’ as broadly synonymous despite subtle differences in their use across disciplinary traditions. However, as we critique the use of those terms, we suggest (and adopt) the use of ‘anomalous’ as a less prejudicial and more inclusive alternative.

The paper is informed by what Daley (2025) and Zumbo (2023) describe as sources of human and contextual variation and by the potential impacts those sources may have on the diverse ways people engage with and respond to tests and assessments. That perspective helps us to expand the discussion of aberrant or deviant responses beyond test disengagement, cheating, and data fabrication. In doing that, we consider what Daley (2025) describes as a primary tenet of universal design: to ‘expect and design for human variation’ (p. 5). In this case, ‘design’ concerns diversity-oriented response models and how we interpret and explain anomalous and unexpected variation.

As Hubley and Zumbo (2017) describe them, response processes are ‘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 score variation’ (Hubley & Zumbo, 2017, p. 2). Their definition ‘expands response processes beyond the cognitive realm, to include emotions, motivations and behaviours’ (p. 2).

As we shall illustrate, that perspective is informed by anthropological and psychometric studies that have captured how assessment responses are mediated by cultural, affective, embodied, ecological and interactive dimensions of testing situations (Zumbo, 2023, Sections 4.1–4.2). The interpretation and explanation of anomalous responses, therefore, requires more than task decomposition and the construction of expected steps and reaction times to consider the multiple factors that influence and mediate response processes – rather than as if tracing the expected sequence of actions, steps, and intervals, that lead (or do not lead) the respondent seamlessly from task or stimulus presentation to response (Zumbo et al., 2023).

While we recognise that anomaly detection can provide an important source of information on threats to test integrity, we argue that its potential as a source of formative insight is hobbled by that negative characterisation. In this paper, we chart a route out of the conceptual impasse by drawing on Cronbach’s work to question the status of aberrant data as de facto evidence of deviant or pathological responses (or individuals).

We propose less-judgemental explanations for anomalous and unexpected responses in relation to sources of human diversity, test design, content, and administration. Our hope is that by treating anomalous and unexpected responses as a source of formative insight, we can realise the potential described in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) to use response processes as a source of evidence on test validity (Goldhammer et al., 2021; Newton, 2019).

Throughout this paper, the context of use informs the meaning of the terms *response* and *responding*. They are sometimes used interchangeably, such as in ‘*aberrant response*’ and ‘*responding*’. Of course, this use is not always appropriate because an aberrant response is not necessarily due to aberrant responding, or, stated otherwise, to an aberrant response process. What appears to be a typical response can result from an aberrant response process, i.e. from aberrant responding such as rapid guessing. The complex relationship between response and responding – and the subtle shift to aberrant *responders* is an example of a situation where the concept of ‘aberrant’ is not particularly useful or meaningful because the relationship between response and responding is intricate, entangled, and difficult to untangle conceptually. This Gordian knot of response, response processes, and respondent is another instance where the aberrant notion is not very informative. Categorising certain response behaviours as aberrant oversimplifies the issue and may reduce our limited understanding of response processes.

Humans in the loop

While a review informs this argument about historical literature, its contemporary significance concerns how we think about, model, and design for diversity in computational and AI-based systems. Recent literature has promoted the idea that ensuring fairness, trust, and integrity in AI-based assessment systems requires keeping humans ‘in the loop’ (Belzak et al., 2026; Hughes & Clesham, 2024). While it is important to recognise the importance of human oversight in AI systems, our argument concerns the human character of judgements of aberrant responding. That is, sources of unfairness and bias did not begin with the invention of Large Language Models (Nisbet & Shaw, 2019).

As Hacking (1983, 1986, 1992, 1995) noted, the ways we humans categorise, interpret, and use data on human behaviour have consequences not only for those individuals but also for our understanding of social and biological phenomena. Hacking’s insights are relevant, as the categories and techniques we use to detect and explain apparently aberrant response processes, responses, and respondents have consequences for test takers (for example, who might be accused of cheating or lacking engagement) and test developers and users who recommend removing a test item, and potentially lead to construct under-representation (Messick, 1989).

Two aspects of Hacking (1983, 1986, 1992, 1995) work are especially relevant to this paper. Hacking argues that data are not neutral reflections of reality but constructed artefacts – marks produced by human judgement and techniques. Regarding response processes research, these marks are interpreted and transformed to infer traits or classify individuals.

Hacking distinguishes raw data from processed forms, emphasising that data gain meaning through deliberate manipulation and modelling. Thus, moving from sensor recordings to traces and patterns of response characteristics – and ultimately to claims about response processes – involves human practices reflecting a purposeful, premeditated, and highly technological research tradition that can take us far away from the raw data. Furthermore, his notion of categorisation as ways of ‘Making Up People’ highlights how scientific classifications, especially in the human sciences, do more than describe – they shape identities and influence behaviour through what he calls the ‘looping effect’.

This concept is not about whether classifications are accurate but rather that social and scientific classifications are not simply descriptive but also prescriptive, shaping how we understand and interact with the world and the people within it. Hacking's (1995) arguments on looping effects inform our argument as we consider how diversity in assessment responding is interpreted and categorised, and its potential consequences. As we will highlight about Cronbach's (1946, 1957) discussion of response sets in elicited responses, questions concerning the detection and interpretation of aberrant *responses* (and item response modelling) have been a source of enquiry and debate for many years (e.g. Levine & Rubin, 1979; Meijer, 1996; Panayides & Tymms, 2013; Van Der Flier, 1982; Wright & Stone, 1979). However, the contemporary use of granular data on response processes in digital and AI-based systems creates additional challenges and responsibilities regarding the opaque nature of process constructs, data interpretation, ethics, and consequences (Siddiq & Murchan, 2024).

Measures of digital process data may be sensitive to sources of diversity in human populations or testing situations in ways that test designers and researchers do not anticipate and that may have unanticipated negative consequences, both for the interpretation and use of data and the potential for Hacking's 'looping effects' to impact how we categorise anomalous and unexpected responses and respondents. As we suggest, the identification and explanation of aberrant response processes require the development of process constructs and response-process models that are attentive to these demands (Bulut, Beiting-Parrish, et al., 2024; Siddiq & Murchan, 2024).

Aberrant responding

The idea of aberrant responses, as we understand it, typically refers to the identification of evidence on test-taker responses, or response processes, that suggests a poor fit with existing response models or that in some way is inconsistent with the assumptions and expectations of test design or test use, and that may indicate a threat to the integrity of assessment data. The concept of aberrance operates at the socio-technical interface between procedures of detection and moral aspects of judgement. In the West, the concept of aberrance has religious, Latin origins, to stray, or deviate from a correct or moral path. As a result, it is sometimes difficult to disentangle the technical and moral aspects of aberrance and judgement. However, those connotations do not necessarily extend globally, suggesting that the ways aberrance is interpreted and treated might be considered from an intercultural perspective. As a result, while the 'detection' of aberrant responses in assessment might be viewed as a statistical procedure, it is not purely a statistical concern. It involves subjective judgements about the legitimate sources of variation in response processes. How such data are identified, interpreted, and explained has consequences for test design, perceptions of test integrity, and the interpretation and use of test scores. Judgements about aberrant data may have negative consequences for those groups and individuals whose responses are deemed aberrant, as well as for the interpretation and use of test data.

The methods used for identifying aberrant data have historically involved statistical procedures to analyse patterns of responses to questionnaires and assessments, and their relationships with test scores and item parameters (e.g. Cronbach, 1946; Lord, 1983). However, with the adoption of digital assessments, the information used to detect

aberrant responses has expanded to include diverse sources of data on *response processes* – in particular, involving inferences about item response times and clickstream (e.g. Bulut et al., 2025; Bulut, Beiting-Parrish, et al., 2024; Jiang et al., 2024; Nagy et al., 2023; Ulitzsch, He, et al., 2024; Wise & Gao, 2017). Those new techniques mark an expanded focus, from the analysis of elicited responses also to including response processes – and by implication, a transition (not always acknowledged and rarely validated) from product (i.e. response-oriented) to process-oriented constructs, with new requirements for validation (Goldhammer et al., 2021; Zumbo et al., 2023).

The use of process-oriented data to analyse anomalous patterns of item responding is likely to expand further with new developments in AI-supported techniques involving immersive, interactive and augmented reality. Anticipating future assessment applications of Generative AI (Jiang et al., 2024; Yan et al., 2024) and their routine analysis of process data, we are concerned with ensuring that those systems use models and terms that recognise legitimate sources of diversity and that their automated data classifications and tags do not over-fit, or misclassify everyday sources of diversity in ways that create ‘systemic biases’ (Yan et al., 2024, p. 10), or that create unintended negative consequences (Hubley & Zumbo, 2011; Messick, 1989). As a result, we need to critically examine the notion of aberrant responding and its role in assessment theory.

Cronbach’s response sets

Cronbach’s (1946) influential paper on response sets identified various tendencies in item responding that can undermine test reliability (e.g. caution, acquiescence, guessing, rushing, evasiveness). Cronbach defined response sets as ‘*Any tendency causing a person to consistently give different responses to test items than he would when the same content is presented in a different form*’ (p. 476). He suggested that the explanation for the presence of response sets could be found in the relationship between those tendencies, which he associated with individual differences, and test item design and content. While Cronbach (1946) argued that ‘response sets should be eliminated where possible’ (p. 487), it was the ‘*factors*’ in test design and administration that he sought to ‘*root out*’ (p. 488) rather than respondents.

Cronbach concluded his 1946 paper by suggesting that further research was required on the ‘nature and origin’ of individual differences observed in response sets. His emphasis on tendencies is inclusive, and he does not use terms such as ‘aberrant’ in his discussion. In his later work (Cronbach, 1957, 1975), he encouraged researchers to accept and investigate unintended sources of variation that might be attributed to individual differences or to variation in the testing situation rather than viewing them as embarrassing or annoying sources of variation that should be ‘*cast into the outer darkness, known as error variance*’ (Cronbach, 1957, p. 674).

Cronbach advocated contextualised and situation-specific studies of unintended sources of variation in response processes. Writing in 1975, he reflected on the need for such investigations: ‘*An observer collecting data in one particular situation is in a position to appraise a practice or a proposition in that setting, observing effects in context*’ (...) ‘*as he goes from situation to situation, his first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that locale of series of events*’ (see Geertz, 1973, Chapter 1, on thick description). *As results accumulate,*

a person who seeks understanding will do his best to trace how the uncontrolled factors could have caused local departures from the modal effect’ (Cronbach, 1975, pp. 124–125).

Our argument on the interpretation of anomalous responding is informed by the kind of ethnographic-psychometric insights that Cronbach described. For over a decade, we have integrated psychometric and anthropological perspectives and methods to explore sources of variation in assessment response processes that are not reflected in psychometric data (Maddox, 2014). Cronbach (1975) suggested that ethnographic descriptions might provide insights into the conditions in testing situations that shape response processes and test scores. Our ethnographic-psychometric study of the UNESCO LAMP assessment in Mongolia used the approach suggested by Cronbach, as Maddox went ‘from situation to situation’ to observe response processes in context (Maddox, 2014, 2015, 2017). Maddox’s subsequent study of the OECD PIAAC assessment in Slovenia used ethnographic methods to observe and explain departures in response processes and expected test protocols (Maddox, 2018a, 2018b; Maddox & Zumbo, 2017; Maddox et al., 2019).

Those studies adopted a non-judgemental, relativist stance to observe sources of variation without prejudice as a source of insight into test performance. In contrast, we find a tradition that has consistently attributed aberrant responses to sources of abnormal pathology and ‘deviant response tendencies’ (Berg, 1957, p. 157).

Abnormal psychology

Berg’s (1957) paper on the ‘Deviation Hypothesis’ presented a provocative critique of Cronbach’s work, portraying aberrant respondents as a problem. Response sets are presented as reflections of ‘patterns of abnormality’, which Berg (1957) attributed (in the parlance of the time in abnormal psychology) to perceived social deviance and pathology – including ‘*schizophrenia, homosexuality, fascist personality, and juvenile delinquency*’. Rejecting Cronbach’s (1957) argument, Berg (1957) claimed, ‘*test item content isn’t important; the deviant responses are*’. (p. 158).

As disturbing as Berg’s (1957) paper might seem to a contemporary reader, we see continuities between Berg’s argument and the wider literature on aberrant responding. For example, Wright and Stone’s (1979) *Best Test Design* (1979), a handbook of Rasch Measurement, described ‘*misfitting response patterns*’ as ‘*sleeping or fumbling*’, ‘*guessing*’ and ‘*plodding*’ (Wright & Stone, 1979, p. 171). They (Wright and Stone) do not seem too concerned with the methodological precision or ethical implications of their conjectures. ‘*Were they sleeping on the easy portion of the test?*’ (Wright and Stone, p. 170). Their approach works towards abstract generalisation rather than ‘contextualised explanation’ (Zumbo, 2009, 2023).

The contemporary research literature on aberrant responding appears to have more in common with Berg, Wright and Stone than with Cronbach, despite Cronbach’s hugely influential status as a psychometrician and educational psychologist. Indeed, Cronbach is rarely cited in contemporary literature on aberrant responding. For example, Meijer’s (1996) discussion of person-misfit and aberrant responding treats Wright and Stone’s (1979) categories as self-evident. While recent literature offers comparatively more elaborate accounts, it continues to frame aberrant responding as a form of deficit or pathology – for example, describing aberrant responders in terms such as *careless*,

cheating, dishonest, disengaged, faking, fraudulent, idle, inattentive, insufficient effort, guessing, lazy, mischievous, non-effortful, non-genuine, plodding, sleeping, and unmotivated. While specific papers do not use all those terms together, their citations and terminology demonstrate continuity with earlier approaches (e.g. Bensch et al., 2019; Bulut et al., 2023, 2025; Cao et al., 2024; Denison & Wiernik, 2020; Goldhammer et al., 2021; Gorgun & Bulut, 2022; Holopainen et al., 2024; Jia et al., 2019; Jiang et al., 2024; Pihlajamäki et al., *in press*; Puck, 2024; Uglanova et al., 2025; Ulitzsch et al., 2022; Ulitzsch, Pohl, et al., 2024; Ulitzsch, He, et al., 2024; Vogelsmeier et al., 2025; Winfred et al., 2021; Zhang et al., 2025).

The use of such descriptions of anomalous responses exposes researchers and data users to potential sources of bias, prejudice and mismeasure. They may also breach professional and ethical research standards. Specifically, the *Standards for Educational and Psychological Testing* states that test-taker categories should ‘avoid labels with unnecessarily stigmatising implications’ (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 136). We can readily anticipate that those descriptions of anomalous responding may create unintended negative consequences, particularly if automated, AI-based systems are used to ‘detect’ and label groups of test takers without adequate validation. Further, inferences, automated capture, classification and use of test taker *emotional affect* in AI/ML systems (Zhang et al., 2025) used in education are likely to fall into ‘high risk’ areas that are prohibited by the Artificial Intelligence Act, Regulation (EU) 2024/1689 and as we discuss below, may create potential discrimination against disabled and neurodiverse test takers, in ways that the EU Act also prohibits.

The use of such methodologically prejudiced terms (i.e. as reified and self-evident categories) also risks imprecision in explanation, as it narrows the space and opportunities for interpretation and consideration of alternative hypotheses and interpretations. For example, some literature treats those terms as synonyms or generic placeholders for some underlying phenomena. As Gorgun and Bulut (2022) wrote, ‘*we assemble all these deviant responses as aberrant response behavior*’ (p. 360, emphasis theirs). At the same time, others use composite categories such as ‘*disengaged test taking*’ (Wise, 2017), ‘*rapid guessing*’ (Bulut et al., 2023), and ‘*Careless and insufficient effort responding (C/IER)*’ (Ulitzsch, He, et al., 2024).

As we have argued, treating aberrant respondents as predefined, homogeneous categories no longer seems viable. However, it is deeply embedded in academic theory and practice. That is most clearly illustrated in relation to the recent literature on rapid responding, which appears to serve as a kernel of truth that informs broader arguments about aberrant responses, response processes, and respondents.

The claim is that rapid responses provide a self-evident data signature of test taker ‘disengagement’, ‘carelessness’, ‘guessing’, ‘insufficient effort’ and ‘noneffortful’ (Anghel et al., 2024; Bulut et al., 2023, 2025; Kröhne et al., 2020; Nagy et al., 2021; Ulitzsch, He, et al., 2024; Vogelsmeier et al., 2025; Wise, 2017, 2019; Wise & Kuhfeld, 2020). Competing methodologies for detecting rapid responses offer various ways to parse and sort responses. Despite the apparent elegance of the idea, the phenomenon of rapid responding is more complex and less self-evident than it might appear. Not only are disengaged rapid responses difficult to detect, but using response times alone is not a particularly reliable indicator of disengagement (Embretson, 2023). Disengagement,

guessing, and insufficient effort may occur across the spectrum of response times and are likely to be a more complex and heterogeneous phenomenon than the literature on the detection and explanation of rapid responses acknowledges. For example, respondents might have many reasons for responding rapidly, such as expert knowledge and high engagement, creative responses, and strategic appraisals of item difficulty (Panayides & Tymms, 2013). A high frequency of rapid responses in assessment data might indicate problems with test design, test content, item design, user experience, test-taker fatigue, poor test inclusivity, or problems with test administration (Maddox, 2017, 2018b; Maddox & Zumbo, 2017). As a result, we are obliged to question whether categories such as ‘careless and insufficient responding’ and ‘disengaged rapid responding’ actually exist as distinct social phenomena.

In the first part of this paper, we described and critiqued the notion of aberrant responding and its distinctly negative characterisation of anomalous and unexpected responses, response processes and respondents. Instead, we argued for an expanded space of explanation informed by the definition of response processes provided by Hubley and Zumbo (2017) and insights from our ethnographic-psychometric research. We then drew on Cronbach (1946, 1957) to argue that interpretations of apparently anomalous and unexpected responses should consider and investigate diverse explanations. In the second part, we explore how such an approach might treat unexpected or anomalous responses as a source of formative insight to improve test performance and validity.

Formative insight

As we have argued, there are usually multiple plausible alternative hypotheses and explanations for why test takers do not behave as intended (Bulut, Beiting-Parrish, et al., 2024; Goldhammer et al., 2021). As a result, we reject the idea that anomalous responses and statistical outliers necessarily threaten the integrity of tests and data. We also question the notion that test takers are disengaged, non-effortful, or lazy because they exhibit rapid responses, item skipping, or guessing behaviours. Instead, we want to consider the possibility that anomalous response processes are evidence of legitimate and engaged test-taker strategies, effortful engagement, test-taker autonomy, playfulness and creativity. Identifying anomalous responses, such as DIF, provides a feedback loop on test performance, integrity, and inclusion. That information can provide a valuable source for cycles of item design, pilot studies, and data validation as tests are used for new purposes and across new contexts and populations.

By their very nature, anomalous responses involve departures from expected response models, test-taking behaviour, and strategies due to unanticipated interactions among the person, the test, the testing situation, and its broader ecology (see Zumbo et al., 2023). As a result, as we have argued, judgements about the integrity of assessment data and the appropriateness of response processes require careful, contextualised and pragmatic explanation (Zumbo, 2009). Therefore, we see our task as not confirming or refuting response models (Kane & Mislevy, 2017), but rather using abductive methods, evidence, and theory to piece together and explore plausible and valid explanations.

From a formative perspective, our goal shifts from detection and exclusion in the aberrant paradigm to identifying new insights from anomalous and unexpected

responses that can improve test design, administration, and accessibility. That may also mean adjusting assessment constructs, ‘process constructs’ (Zumbo et al., 2023), thresholds and parameters to increase their tolerance of legitimate diversity and to avoid overfitting. I.e., expanding the range of responses and behaviours that count as normal. We prefer the idea of expanded construct representation (Messick, 1989) in response models and constructs rather than the notion of ‘accommodations’ (i.e. as special exemptions).

We view the potential for deeper, more precise explanations of anomalous and unexpected responses as more effectively achieved through such recognition of test diversity than through developing more efficient methods for detecting and excluding aberrant responses. We see those deeper formative insights into test performance as a more effective route to improved test performance, integrity and inclusion. For example, we can consider the potential of data on anomalous responses to generate new insights into the performance, inclusion, and validity of assessments involving participants with disabilities and neurodiversity. As we have noted, the presence of disability and neurodiversity in tested populations is rarely acknowledged or considered as potential sources of legitimate variation in response processes (Zumbo et al., 2023). Despite systematic attempts to justify the exclusion of such groups from large-scale assessments (Alzahrani, 2024; LeRoy et al., 2018), they are likely to constitute a significant population in most, if not all, assessments and questionnaires.

Diversity and inclusion

Inclusive assessments not only involve designing tests and items to anticipate the diverse needs of test takers but also ensure that the claims made about test scores and performance are independent of individual or group characteristics (i.e. free of bias). They also require supporting evidence on the extent to which such inclusion has been achieved. As a result, the identification of anomalous responses provides an ideal opportunity to develop more inclusive response models and improve inclusive designs that anticipate diversity as a norm (Sireci, 2020). Nevertheless, our literature review found the topic almost entirely ignored in the literature on aberrant responding. See van der Linden and Guo (2008) for a brief mention of dyslexia as a possible source of ‘irregular behaviour’ (p. 365).

There are many ways in which disabilities and neurodiverse conditions, including chronic health conditions such as anxiety and depression, might impact assessment response processes. Those include the potential impacts of embodied and sensory differences, difficulties of self-regulation, fatigue and time management, fine-motor difficulties associated with mouse and cursor movement and tool manipulation, constraints to working memory, reading and information processing, and higher prevalence of anxiety (Astle et al., 2022; Farroni et al., 2022; Robinson, 2025). As a result, we must consider and investigate the possibility that some anomalous patterns in response times and clickstream may reflect the presence of disability and neurodiversity or indicate a lack of test accessibility and inclusion. The risk, otherwise, is of unintended negative consequences, as respondents with disabilities and neurodiverse conditions who are effortful and engaged might be misrepresented as indicating responses that are characterised as evidence of disengagement, carelessness, or insufficient effort.

Discussion

In this paper, following the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), we have sought to refine and improve the treatment of aberrant responses and response processes as a valuable source of validity evidence in psychological and educational assessments. As our review of the contemporary literature has highlighted, researchers have sought to leverage the affordances of digital process data to detect aberrant response processes more reliably. Our argument is not against developing new methods to detect anomalies and unexpected responses in assessment response processes. Indeed, we welcome the development of those new measurement opportunities and techniques as sources of formative insight. However, as we have argued, attempts to detect and categorise anomalous data as *de facto* threats to test integrity narrow the space for interpretation and explanation and, in doing so, may produce a series of unintended negative consequences.

Of those consequences, the first, as we have highlighted, concerns the uniquely negative characterisation of respondents who display anomalous and unexpected responses and response processes. As we have highlighted in our discussion of Cronbach's work on response sets, explanations that attribute aberrant or deviant responses to test-taker traits (i.e. lack of effort) may neglect a host of other potential explanations, such as those of item design, content and administration, and the presence of other factors in the testing situation.

As Markus and Borsboom (2024) stated, 'Zumbo (2023) offers a contemporary development of an approach to test validity centred on explanation influenced by Cronbach's explanatory turn' (p. 10). If so, we might expand the explanation space to include the various dimensions of test design, content, and administration that Cronbach's work highlighted.

As the contrasting arguments of Cronbach (1946, 1957) and Berg (1957) illustrate, a sharp paradigmatic distinction between test-taker traits and characteristics and those of test content, design, and administration is problematic. On either side of that divide, it is likely to neglect important aspects of human diversity as mediating factors in how people engage with assessment tasks that might become evident in data on response processes. As we have suggested, that is illustrated in the complex ways in which neurodiversity, disability, or chronic illness might influence cognitive, affective, developmental, and sensory differences, thereby shaping the ways that test takers experience and engage with test content, designs, and constructs (Zumbo et al., 2023). We see no reason why an explanation-oriented perspective on anomalous responses should not also consider human diversity.

A second set of consequences concerns the 'washback' of 'method' on theory (the hobbling that we alluded to). Methods that integrate the detection of aberrant responses with their classification, for example, such as methods for detecting 'Careless and insufficient effort responding (C/IER)' (Ulitzsch, He, et al., 2024), not only narrow the space of explanation (the term 'detection' condenses identification, categorisation and explanation). They also give an inaccurate impression that the theoretical and empirical work of explanation is complete.

As a result, our suggestion is to procedurally and conceptually disentangle the detection of anomalous responses from their explanation and to critically reflect on the ways in which theoretical assumptions, for example, about test-taker characteristics, test ecologies, and process-oriented test constructs, might tacitly inform detection methods (e.g. the selection of thresholds, parameters, and distributions). In that sense, we can distinguish response times (i.e. as the temporal sum of a sequence of interactions) from the diverse processes and factors that produce those observed outcomes.

These considerations are important as the field transitions from product (sum) related test constructs to those that integrate (either implicitly or explicitly) process-related constructs and models (Goldhammer et al., 2021; Zumbo et al., 2023). Process-oriented constructs, therefore, need to be sufficiently elaborate and multi-dimensional to account for the numerous factors that might legitimately influence variation in item responding as consistent with Hubley and Zumbo's (2017) definition of assessment response processes. An implication is that 'process constructs' (Zumbo et al., 2023), models, and methods must be able to effectively distinguish between legitimate sources of anomalous and unexpected variance that reflect human diversity, and those (process) 'construct irrelevant' (Messick, 1989) response processes and responses that present threats to test integrity. That challenge takes us to the heart of what it means to identify and explain aberrant (or anomalous) response processes.

Conclusions

Advances in the analysis of digital data on response processes in psychological and educational assessments and questionnaires have supported the development of an extensive literature aimed at improving the detection and removal of aberrant responses. As we have discussed in this paper, the literature has usually treated aberrant responses and response processes as a *de facto* threat to test and data integrity, and has treated aberrant respondents through various accusations of social or pathological deviance (i.e. as indicators of insufficient effort, carelessness, or cheating).

We have argued for adopting an alternative paradigm informed by the work of Lee Cronbach. In addition to its questionable ethics (i.e. potentially biased misrepresentations of human diversity), reified categories associated with aberrance, such as 'careless and insufficient', reduce the space for enquiry and explanation. As a result, it limits opportunities to use information on anomalous and unexpected response processes to inform improvements in test performance, test integrity, and inclusion.

We started by reviewing the historical literature on aberrant responses in relation to Cronbach's work on response sets (see Cronbach, 1946, 1957). Cronbach's 'tendencies' (Cronbach, 1946, p. 476) describe a relationship between the test (format, content) and the test taker's responses. In contrast, Berg (1957) portrayed deviant responses as evidence of respondent pathology and social deviance. Like Cronbach, we reject the idea that anomalous responses should be explained that way. As we have argued in this paper, many factors might legitimately produce anomalous response processes and responses (e.g. human diversity, creativity, test ecologies). We therefore suggest replacing the term 'aberrant' with less prejudicial terms such as 'anomalous' or 'unexpected', and

expanding the space of interpretation in ways that anticipate and more rigorously model the sources of human diversity.

As we discussed in relation to the topic of disability and neurodiversity, characterisations of aberrant responses as indications of inattentive, careless responding, or insufficient effort (lazy), respondent disengagement, or insufficient effort are particularly problematic as they have the potential to misrepresent and misunderstand the complex and often difficult experience of test takers with regard to test inclusion and accessibility. Instead, we have argued that methods that identify anomalous and unexpected responses and response processes represent an important source of formative insight that can help inform and refine response models, expand construct representation, and improve test performance, integrity, and inclusion.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Dr. Bryan Maddox is Research Director for Digital Assessment Futures at the Digital Education Futures Initiative, University of Cambridge, and Executive Director of Assessment Micro-Analytics Ltd. Bryan is an Honorary Professor at the University of East Anglia, where he was previously Professor of Educational Assessment. He has held Visiting Professor roles at the Centre for Educational Measurement, University of Oslo, and at the University of British Columbia.

Bruno D. Zumbo is Professor and Distinguished University Scholar at the University of British Columbia, where he holds the Tier 1 Canada Research Chair in Psychometrics and Measurement. His research integrates mathematical analysis, statistics, and philosophies of science to advance theories of measurement, testing, and survey methodology in the social, behavioral, and health sciences. He has held academic appointments at the University of Ottawa and the University of Northern British Columbia, and visiting appointments at University of Massachusetts-Amherst, Harvard University, and University of Cambridge. His contributions span psychometrics, reliability and validity theories, test bias, and the mathematical foundations of test theory.

ORCID

Bryan Maddox  <http://orcid.org/0009-0007-7709-9845>

Bruno D. Zumbo  <http://orcid.org/0000-0003-2885-5724>

References

- Alzahrani, H. A. (2024). Access denied: Assessing the challenges that prevent students with disability to participate in PISA. *The Educational Review, USA*, 8(11), 1284–1295. <https://doi.org/10.26855/er.2024.11.002>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. <https://www.apa.org/pubs/books/standards-testing>

- Anghel, E., Khorramdel, L., & von Davier, M. (2024). The use of process data in large-scale assessments: A literature review. *Large-Scale Assessments in Education*, 12, Article 13. <https://doi.org/10.1186/s40536-024-00202-1>
- Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual research review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, 63(4), 397–417. <https://doi.org/10.1111/jcpp.13481>
- Belzak, W., Baig, B., Cardwell, R., LaFlair, G., Liao, M., Niu, C., Sobczack, M., Socias, C., & Shih, Y. S. (2026). *Duolingo English test: Security and test integrity* (Duolingo research report). Duolingo.
- Belzak, W., Burstein, J., & von Davier, A. A. (2025). Evaluating fairness in AI-assisted remote proctoring. *Proceedings of Machine Learning Research*, 273, 125–132. <https://proceedings.mlr.press/v273/belzak25a.html>
- Bensch, D., Maaß, U., Greiff, S., Horstmann, K. T., & Ziegler, M. (2019). The nature of faking: A homogeneous and predictable construct? *Psychological Assessment*, 31(4), 532–544. <https://doi.org/10.1037/pas0000619>
- Berg, I. A. (1957). Deviant responses and deviant people: The formulation of the deviation hypothesis. *Journal of Counseling Psychology*, 4(2), 154–161.
- Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, 5(3), Article 3. <https://doi.org/10.59863/MIQL7785>
- Bulut, O., Gorgun, G., & He, S. (2024). Unsupervised anomaly detection in sequential process data: Insights from PIAAC problem-solving tasks. *Zeitschrift für Psychologie*, 232(2), 74–94. <https://doi.org/10.1027/2151-2604/a000558>
- Bulut, O., Gorgun, G., & Karamese, H. (2025). Incorporating test-taking engagement into multi-stage adaptive testing design for large-scale assessments. *Journal of Educational Measurement*, 62(1), 57–80. <https://doi.org/10.1111/jedm.12380>
- Bulut, O., Gorgun, G., Wongvorachan, T., & Tan, B. (2023). Rapid guessing in low-stakes assessments: Finding the optimal response time threshold with random search and genetic algorithm. *Algorithms*, 16(2), Article 89. <https://doi.org/10.3390/a16020089>
- Cao, N., Finos, L., Lombardi, L., & Calcagni, A. (2024). A novel cfa+efa model to detect aberrant respondents. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(5), 1283–1309. <https://doi.org/10.1093/jrsssc/qlae036>
- Cronbach, L. J. (1946). Response sets and test validity. *Educational and Psychological Measurement*, 6(4), 475–494. <https://doi.org/10.1177/001316444600600405>
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *The American Psychologist*, 12(11), 671–684. <https://doi.org/10.1037/h0043943>
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *The American Psychologist*, 30(2), 116–127. <https://doi.org/10.1037/h0076829>
- Daley, S. G. (2025). Universal design of educational psychology? Improving theory and application by focusing on students with disabilities. *Educational Psychologist*, 60(3), 141–153. <https://doi.org/10.1080/00461520.2024.2441661>
- Denison, A. J., & Wiernik, B. M. (2020, December 18). Careless response processes are heterogeneous: Comment on Goldammer, et al. <https://doi.org/10.31234/osf.io/9rk4w>
- Embretson, S. (2023). Understanding examinees' item responses through cognitive modeling of response accuracy and response times. *Large-Scale Assessments in Education*, 11(1), Article 9. <https://doi.org/10.1186/s40536-023-00158-8>
- Farroni, T., Valori, I., & Carnevali, L. (2022). Multimedia interventions for neurodiversity: Leveraging insights from developmental cognitive neuroscience to build an innovative practice. *Brain Sciences*, 12(2), 147. <https://doi.org/10.3390/brainsci12020147>
- 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), 20. <https://doi.org/10.1186/s40536-021-00113-5>

- Gorgun, G., & Bulut, O. (2022). Identifying aberrant responses in intelligent tutoring systems: An application of anomaly detection methods. *Psychological Test and Assessment Modeling*, 64(4), 359–384.
- Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.
- Hacking, I. (1986). Making up people. In T. C. Heller, M. Sosna, & D. E. Wellbery (Eds.), *Reconstructing individualism: Autonomy, individuality, and the self in Western thought* (pp. 222–236). Stanford University Press.
- Hacking, I. (1992). The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Science as practice and culture* (pp. 29–64). University of Chicago Press.
- Hacking, I. (1995). The looping effects of human kinds. In D. Sperber & J. Premack (Eds.), *Causal cognition* (pp. 351–383). Oxford University Press.
- Holopainen, S., Metsämuuronen, J., Laakso, M.-J., & Kujala, J. (2024). Mutual information: A robust and valid tool to identify rapid-guessing behavior. [Preprint]. ResearchGate. <https://doi.org/10.13140/RG.2.2.23818.07368>
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research*, 103(2), 219–230. <https://doi.org/10.1007/s11205-011-9843-4>
- Hubley, A. M., & Zumbo, B. D. (2017). Response processes in the context of validity: Setting the stage. In B. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer International Publishing/Springer Nature. https://doi.org/10.1007/978-3-319-56129-5_1
- Hughes, S., & Clesham, R. (2024). *Test integrity: Ensuring trust by integrating human judgment and AI systems*. Pearson. <https://www.pearsonpte.com/content/dam/ELL/pte/pearsonpte/general/test-integrity-ensuring-trust-by-integrating-human-judgement-and-ai-systems-november-2024.pdf>
- Jia, B., Zhang, X., & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. *Frontiers in Psychology*, 10, Article 43. <https://doi.org/10.3389/fpsyg.2019.00043>
- Jiang, Y., Zhang, M., Hao, J., Deane, P., & Li, C. (2024). Using keystroke behavior patterns to detect nonauthentic texts in writing assessments: Evaluating the fairness of predictive models. *Journal of Educational Measurement*, 61(4), 571–594. <https://doi.org/10.1111/jedm.12416>
- Kane, M. T., & Mislevy, R. (2017). Validating score interpretations based on response processes. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 11–24). Routledge.
- Kröhne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, 62(2), 147–177.
- LeRoy, B. W., Samuel, P., Deluca, M., & Evans, P. (2018). Students with special educational needs within PISA. In *Assessment in education: Principles, policy & practice*. Taylor and Francis.
- Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4(4), 269–290. <https://doi.org/10.3102/10769986004004269>
- Lord, F. (1983). Maximum likelihood estimation of item response parameters when some responses are omitted. *Psychometrika*, 48(3), 477–482. <https://doi.org/10.1007/BF02293689>
- Maddox, B. (2014). Globalising assessment: An ethnography of literacy assessment, camels and fast food in the Mongolian Gobi. *Comparative Education*, 50(4), 474–489. <https://doi.org/10.1080/03050068.2013.871440>
- Maddox, B. (2015). The neglected situation: Assessment performance and interaction in context. *Assessment in Education: Principles, Policy & Practice*, 22(4), 427–443. <https://doi.org/10.1080/0969594X.2015.1026246>
- 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. (2018a). Ethnography and assessment at the bottom of the pyramid. In D. A. Wagner, S. Wolf, & R. F. Boruch (Eds.), *Learning at the bottom of the pyramid: Science, measurement and policy in low-income countries* (pp. 141–153). IIEP.
- Maddox, B. (2018b). Interviewer-respondent interaction and rapport in PIAAC. *Quality Assurance in Education*, 26(2), 182–195. <https://doi.org/10.1108/QAE-05-2017-0022>

- Maddox, B., Keslair, F., & Javrh, P. (2019). Investigating testing situations. In B. Maddox (Ed.), *International large-scale assessments in education: Insider research perspectives* (pp. 157–172). Bloomsbury Academic.
- Maddox, B., & Zumbo, B. D. (2017). Observing testing situations: Validation as jazz. In B. D. Zumbo & A. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 179–192). Springer/Nature. https://doi.org/10.1007/978-3-319-56129-5_1/
- Markus, K. A., & Borsboom, D. (2024). *Frontiers of test validity theory: Measurement, causation, and meaning* (2nd ed.). Routledge.
- Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3–8. https://doi.org/10.1207/s15324818ame0901_2
- 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>
- Nagy, G., Ulitzsch, E., & Lindner, M. A. (2023). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, 39(3), 751–766. <https://doi.org/10.1111/jcal.12719>
- Nagy, G., Ulitzsch, E., & Lindner, M. A. (2023). The role of rapid guessing and test-taking persistence in modelling test-taking engagement. *Journal of Computer Assisted Learning*, 39(3), 751–766.
- Newton, P. E. (2019). What is response process validation evidence and how important is it? A paper 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>
- Nisbet, I., & Shaw, S. (2019). Fair assessment viewed through the lenses of measurement theory. *Assessment in Education: Principles, Policy & Practice*, 26(5), 612–629. <https://doi.org/10.1080/0969594X.2019.1586643>
- Panayides, P., & Tymms, P. (2013). Investigating whether aberrant response behaviour in classroom maths tests is a stable characteristic of students. *Assessment in Education: Principles, Policy & Practice*, 20(3), 349–368. <https://doi.org/10.1080/0969594X.2012.723610>
- Pihlajamäki, M., Eisele, G., Peeters, L., Kirtley, O. J., Myin-Germeys, I., & Lafit, G. (in press). An investigation of the temporal dynamics of careless responding across different populations in experience sampling data. *Assessment*. <https://doi.org/10.1177/10731911251410340>
- Puck, M. R. (2024). *Non-genuine responses and how to find them* [Master's thesis, Centre for Educational Measurement, University of Oslo].
- Robinson, K. F. (2025). Introduction. In K. F. Robinson, M. T. Carew, & N. E. Grace (Eds.), *Inaccessible access: Rethinking disability inclusion in academic knowledge creation* (pp. 1–18). Rutgers University Press.
- Siddiq, F., & Murchan, D. (2024). Towards a code of ethics for using technology-enabled data and related analytic approaches in educational assessment. *Assessment in Education: Principles, Policy & Practice*, 31(5–6), 376–400. <https://doi.org/10.1080/0969594X.2025.2453138>
- Sireci, S. G. (2020). Standardisation and “understand” arization in educational assessment. *Educational Measurement: Issues and Practice*, 39(3), 100–105. <https://doi.org/10.1111/emip.12377>
- Uglanova, I., Nagy, G., & Ulitzsch, E. (2025). A mixture IRT model for handling different types of careless respondents [Preprint]. OSF. <https://osf.io/mzb3e/>
- Ulitzsch, E., He, Q., & Pohl, S. (2022). Using sequence mining techniques for understanding incorrect behavioral patterns on interactive tasks. *Journal of Educational and Behavioral Statistics*, 47(1), 3–35. <https://doi.org/10.3102/10769986211010467>
- Ulitzsch, E., He, Q., & Pohl, S. (2024). Innovations in exploring sequential process data. *Zeitschrift für Psychologie*, 232(2), 71–73. <https://doi.org/10.1027/2151-2604/a000560>
- Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2024). Using response times for joint modeling of careless responding and attentive response styles. *Journal of Educational and Behavioral Statistics*, 49(2), 173–206.
- Van Der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology*, 13(3), 267–298. <https://doi.org/10.1177/0022002182013003001>

- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384. <https://doi.org/10.1007/s11336-007-9046-8>
- Vogelsmeier, L. V. D., Uglanova, I., Rein, M. T., & Ulitzsch, E. (2025). Investigating dynamics in attentive and inattentive responding together with their contextual correlates using a novel mixture IRT model for intensive longitudinal data. *British Journal of Mathematical and Statistical*, 79(2), 379–408.
- Winfred, A., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology & Organizational Behavior*, 8, 105–137.
- Wise, S. L. (2017). Rapid-guessing behavior: ‘Its identification, interpretation, and implication’. *Educational Measurement: Issues and Practice*, 36(4), 52–61.
- Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education*, 32 (4), 325–336.
- 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. (2020). A cessation of measurement: Identifying test taker disengagement using response time. In M. J. Margolis & R. A. Feinberg (Eds.), *Integrating timing considerations to improve testing practises* (pp. 150–164). Routledge.
- Wright, D., & Stone, M. H. (1979). *Best test design*. Mesa Press.
- Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. *Nature Human Behaviour*, 8(10), 1839–1850. <https://doi.org/10.1038/s41562-024-02004-5>
- Zhang, L., Domingue, B., Vogelsmeier, L., & Ulitzsch, E. (2025). A beta mixture model for careless respondent detection in visual analogue scale data. OSF. https://doi.org/10.31219/osf.io/tp6df_v1
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233. <https://doi.org/10.1080/15434300701375832>
- Zumbo, B. D. (2009). Validation as contextualised 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). Information Age Publishing.
- Zumbo, B. D. (2023). A dialectic on validity: Explanation-focused and the many ways of being human. *International Journal of Assessment Tools in Education*, 10(Special Issue), 1–96. <https://doi.org/10.21449/ijate.1406304>
- Zumbo, B. D., Maddox, B., & Care, N. M. (2023). Process and product in computer-based assessments: Clearing the ground for a holistic validity framework. *European Journal of Psychological Assessment*, 39(4), 252–262. <https://doi.org/10.1027/1015-5759/a000748>