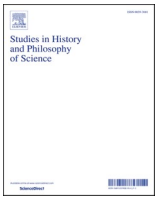
Studies in History and Philosophy of Science 116 (2026) 102123

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa



Rethinking psychological measurement: Validity potential versus realised validity

Wendy C. Higgins [a,b,*], David M. Kaplan [b,c,e], Alexander J. Gillett [d], John Sutton [f], Robert M. Ross [d]

[a] University of Melbourne, Melbourne School of Psychological Sciences, VIC, 3010, Australia
[b] Macquarie University, School of Psychological Sciences, NSW, 2109, Australia
[c] Macquarie University, Performance and Expertise Research Centre, NSW, 2109, Australia
[d] Macquarie University, School of Humanities, NSW, 2109, Australia
[e] Macquarie University, Minds and Intelligences Research Centre, NSW 2109, Australia
[f] University of Stirling, Centre for the Sciences of Place and Memory, Stirling, FK9 4LA, Scotland, UK

A B S T R A C T

We propose a concept of validity with a novel feature that we argue can facilitate improved measurement validation practices in the psychological sciences. Following Borsboom and colleagues, our concept of validity is measurement-specific and causal. This contrasts with current guidelines linking validity to the acceptability of both measurement and non-measurement-based interpretations of test scores. Benefits of a measurement-specific concept of validity are that it can make the requirements for valid measurement clearer and make validity claims easier to interpret, which we illustrate by comparing the use of test scores for measurement versus prediction. Our concept of validity also maintains that a causal relationship of sufficient strength from the attribute being measured to the measurement outcomes is necessary and sufficient for valid measurement. This places causal explanations at the centre of the validation process. While causal complexity will make the evaluation of psychological measurements as causal inferences extremely challenging, we describe how the interventionist theory of causation and related work on causal inference can serve as a starting point for addressing this challenge. The novel feature of our concept of validity is that it makes a distinction between the *validity potential* of measurement procedures *in abstracto* (e.g., tests) and the *realised validity* of concrete measurement outcomes (e.g., specific test scores). We describe key benefits of this novel distinction, including its potential to encourage the theoretical refinement of concepts, guide the selection of appropriate measurement procedures for use in research, and increase sample-specific validity evidence reporting.

When citing this paper, please use the full journal title *Studies in History and Philosophy of Science*.

"A theory of validity that leaves one with the feeling that every single concern about psychological testing is relevant, important, and should be addressed in psychological testing cannot offer a sense of direction to the working researcher." (Borsboom et al., 2004, p. 138).

1. Introduction

Psychological research relies on valid psychological measurement.[1] Yet, a growing body of literature demonstrates serious problems with measurement validation practices in the psychological sciences. Of particular concern, review articles across diverse areas of psychological research indicate that validity evidence is infrequently reported, leaving the validity of many psychological measurements uncertain (Alexandrova & Haybron, 2016; Flake & Fried, 2020; Fried et al., 2022; Higgins et al., 2024; Schimmack, 2021; Slaney, 2017).[2] This uncertainty

This article is part of a special issue entitled: Measuring the Human published in Studies in History and Philosophy of Science.

* Corresponding author. Melbourne School of Psychological Sciences, University of Melbourne, Parkville, VIC, 3010, Australia.
E-mail address: wendy.higgins@unimelb.edu.au (W.C. Higgins).

[1] There is little agreement about how best to define measurement (Tal, 2020). We adopt Nunnally's (1978) definition of measurement as "consist[ing] of rules for assigning numbers to objects in such a way as to represent quantities of attributes" (p. 3).

[2] "Validity" is an ambiguous term. In the context of psychological research, it can refer to the validity of measurements (i.e., "construct validity"), the validity of statistical conclusions (i.e., "statistical-conclusion validity"), the validity of causal inferences made within a study (i.e., "internal validity"), and the validity of generalisations of causal inferences beyond a study (i.e., "external validity"; Shadish et al., 2002; Vazire et al., 2022). We use the term "validity" in relation to measurement.

https://doi.org/10.1016/j.shpsa.2026.102123
Received 29 June 2024; Received in revised form 14 October 2025; Accepted 14 January 2026

Fig. 1. One of 36 items from the Eyes Test. Participants select which of the four mental states best matches what a person is thinking or feeling. The target response is "distrustful" (Baron-Cohen et al., 2001). The Eyes Test is freely available from https://www.autismresearchcentre.com/tests/.

has profound implications not only for the conclusions of individual psychological research studies but also for theoretical and applied research that builds on those studies and applications of those studies' findings.

It has been argued that an overly broad concept of validity and a resulting lack of clarity about how researchers should assess validity are key contributors to poor measurement validation practices in the psychological sciences (e.g., Borsboom et al., 2004; Cizek, 2012). To address this issue, Borsboom and colleagues (2004) proposed a concept of validity that they argue is "simpler … [and] theoretically superior to the position taken in the existing literature" (p. 1061):

> "[A] test is valid for measuring an attribute if and only if (a) the attribute exists and (b) variations in the attribute causally produce variations in the outcomes of the measurement procedure." (Borsboom et al., 2004, p. 1061)

Borsboom and colleagues' (2004) concept of validity has two key features that make it well-suited to guiding psychological measurement validation practices: it focuses exclusively on measurement, and it explicitly stipulates a causal relationship from attributes to measurement outcomes as a necessary condition for valid measurement. Nonetheless, their concept of validity has three important limitations in the context of psychological research. First, it makes realism about psychological attributes a requirement for valid measurement, which is unnecessarily restrictive (Eronen, 2025). Second, as currently formulated, the necessary and sufficient causal condition is too inclusive (Eronen, 2025). Third, ascribing validity to tests may draw attention away from the impact that contextual factors have on validity (Larroulet Phillipi, 2021).

In this paper, we propose an alternative concept of validity that addresses the limitations of Borsboom and colleagues' (2024) concept of validity while retaining its two key benefits. To illustrate the advantages of our proposed concept of validity, we use examples based on the "Reading the Mind in the Eyes" Test (hereafter the Eyes Test; Baron-Cohen et al., 2001). The Eyes Test is a multiple-choice test that asks participants to select which of four mental states best matches what a person is thinking or feeling based on a photograph of a portion of their face that includes the eyes (see Fig. 1). It is one of the most widely used measures of adult social cognitive ability (Yeung et al., 2024),[3] and it comes highly recommended, being listed in the United States of America's National Institute of Mental Health's (NIMH) Research Domain Criteria Initiative as a "current best option" for a task to assess the "perception and understanding of others" (NIMH, 2016). The Eyes Test is also frequently referred to as a "valid" or "validated" test of social cognitive ability (Higgins et al., 2024).

Despite being widely used and widely accepted as one of the best measures of social cognitive ability currently available, research has demonstrated that the validity evidence for the Eyes Test is inadequate (Betz et al., 2019; Higgins et al., 2024; Higgins, Ross, et al., 2023; Higgins, Savalei, et al., 2023, 2025; Johnston et al., 2008; Silverman, 2022). Particularly comprehensive evidence comes from Higgins and colleagues (2024) who surveyed 1,461 studies that administered the Eyes Test and found that very few studies reported six key types of quantitative validity evidence (test-retest reliability: 2% of surveyed studies; internal consistency: 11%; factor structure: <1%; known groups: 3%; convergent: 11%; and discriminant: 3%).[4] Moreover, the validity evidence that was reported frequently indicated poor validity, and follow up work by Higgins and colleagues (Higgins, Kaplan, et al., 2025) highlighted two serious flaws with the Eyes Test: it is implausible that people can reliably identify complex mental states from the static, decontextualised, visual stimuli used in the Eyes Test, and the "correct" responses are not objectively correct. These findings, we argue, are inconsistent with the Eyes Test being a "validated" measure of social cognitive ability. Thus, the Eyes Test provides a striking example of a failure of measurement validation practices: a test with inadequate validity evidence has been frequently used, widely reported to be "valid", and recommended as a current best option to measure social cognitive ability.

Given the increasing body of literature demonstrating inadequate measurement validation practices across the psychological sciences (e.g., Flake & Fried, 2020; Slaney, 2017), the voluminous Eyes Test literature appears to be symptomatic of a widespread credulity toward the validity of psychological measurements, despite insufficient evidence. Following Borsboom and colleagues (2004), we contend that a suitable concept of validity is crucial for reducing this credulity and improving measurement validation practices. A key benefit of our proposed concept of validity is that it can readily translate into concrete guidelines for practicing researchers, which, in turn, can facilitate greater critical assessment of psychological measurements. This serves our broader aim of increasing the reliability of psychological research.

In Section 2, we describe the benefits of a measurement-specific concept of validity relative to the broader concept currently

[3] The social cognition literature is replete with overlapping constructs, and it is not always clear when terms are used synonymously or to refer to different constructs. Higgins and colleagues (2024) identified more than 50 different terms describing what the Eyes Test measures. Consequently, we use "social cognitive ability" as an umbrella term.

[4] We consider reliability to be necessary, but not sufficient, for valid measurement (McNeish, 2024; Slaney, 2017). Higher levels of reliability indicate lower levels of measurement error, and test scores cannot be valid without enough reliability to support the interpretation that variance in test scores represents variance in the construct of interest rather than resulting from measurement error. Therefore, evidence of reliability, including test-retest reliability and internal consistency, are important sources of evidence for measurement validity. However, reliability is not sufficient for valid measurement because it does not tell us *what* is being measured, and test scores can be reliable without being valid measurements of the target psychological construct (see Section 5).

underpinning key validation guidelines. In Section 3, we propose a necessary and, in principle, sufficient causal condition for valid measurement that builds on the work of Borsboom and colleagues (2004) and Eronen (2025). In Section 4, we argue that a distinction exists between what we refer to as the *validity potential* of measurement procedures *in abstracto* (e.g., tests) and the *realised validity* of measurement outcomes obtained from each specific instantiation of those procedures (e.g., test scores collected in a study),[5] and we describe four benefits of making this distinction explicit in the context of psychological research. Finally, in Section 5, we discuss how the interventionist theory of causality (Pearl, 2009; Woodward, 2003) and methods for modelling causal relationships (e.g., Bulbulia, 2024; Rohrer, 2024) could offer a useful starting point for evaluating measurement procedures as causal relationships and validating psychological measurements as causal inferences. We also consider how existing sources of validity evidence could be interpreted in relation to causal models of measurement procedures.

2. The case for a measurement-specific concept of validity

One widespread and intuitive way to conceptualise validity is to ascribe validity to tests[6] such that a valid psychological test is one that measures what it is intended to measure (Borsboom et al., 2004; Eronen, 2025; Kelley, 1927). By contrast, several influential construct validity theorists ascribe validity to *interpretations* of test scores (e.g., Messick, 1989; Clark & Watson, 2019), which may or may not relate to measurement (Hood, 2009). The American Psychological Association (APA), which is the largest professional and scientific organisation of psychologists in the United States of America, adopts the latter position: "the concept of validity is comprehensive and refers not only to test characteristics but also to the appropriateness of test use and to the accuracy of the inferences made on the basis of test scores" (Sireci & Sukin, 2013, p. 61; see also American Education Research Association [AERA] et al., 2014; Appelbaum et al., 2018). Moreover, key APA texts explicitly state that validity should be ascribed to interpretations of test scores not to tests (see Table 1).

These two conceptions of validity differ in a number of ways. As already mentioned, they differ in scope (i.e., restricted to measurement versus including various interpretations and uses of test scores); and they differ in where validity is ascribed (i.e., to tests versus interpretations of test scores; Hood, 2009). Other researchers have noted that, while Borsboom and colleagues' (2004) focus is on ontology and the concept of validity itself, Messick (1989) and key APA texts are concerned with semantics, epistemology, and the validation process (Borsboom et al., 2004; Hood, 2009; Larroulet Philippi, 2021). Here we focus on the difference in scope. Whether validity is restricted to measurement or not has important implications for how we define validity *and* how we validate psychological measurements. The appropriateness and justifiability of non-measurement-based interpretations and uses of test scores (e.g., prediction and diagnosis) are unquestionably important considerations for psychological test use. Nonetheless, having a concept of validity that is explicitly restricted to measurement, while other, non-measurement related interpretations and uses of test scores are assessed separately, can benefit psychological research in at least two important ways.

First, a measurement-specific concept of validity can make ascriptions of validity easier to interpret and avoid misinterpretations of validity claims. This is because, under a measurement-specific concept of validity, every validity claim pertains to what a test does (or does not) measure. By contrast, under a concept of validity that includes other inferences based on test scores and the appropriateness of test uses, an ascription of validity does not necessarily provide *any* information about what a test measures. This is because neither the accuracy of inferences based on test scores nor the appropriateness of uses of tests are necessarily reliant on what a test measures.

For illustrative purposes, we focus on the distinction between the use of test scores for measurement versus prediction, which are two of the most prominent uses of psychological test scores (Smits et al., 2018). Consider an example from the *Standards for Educational and Psychological Testing* (Hereafter "the *Standards*"; AERA et al., 2014), which are endorsed by the APA and considered to be a key reference for best practices in the validation of psychological measurements (e.g., Cizek, 2020; Flake et al., 2017). The *Standards* describe how scores on a test of mathematics achievement (what the test measures) could be interpreted as predicting student success at college-level work (inference based on test scores) for the purpose of informing college admission decisions (test use). Crucially, in this example, neither the accuracy of the prediction nor the appropriateness of the test use requires that the test measures mathematics achievement. For instance, if the test contains word problems, it might primarily function as a measure of verbal ability rather than mathematics achievement (Strohmaier et al., 2022). Nonetheless, if verbal ability is an accurate predictor of college-level success, then the test could accurately predict college performance despite being a poor measure of mathematics achievement. Moreover, if validity is ascribed based on the appropriateness of the intended interpretation of test scores, validity can obtain in cases like this, irrespective of what a test measures. Conversely, even if the test is a very accurate measure of mathematics achievement, under the broad concept of validity, validity should not be ascribed in this case if the test is a poor predictor of college-level success. Thus, as this example illustrates, ascriptions of validity under the broad concept of validity cannot be assumed to tell us anything about how well a test measures what it is supposed to measure. However, if "[v]alidity … as it is understood by almost everybody except construct validity theorists—does patently not refer to a property of test score interpretations, but to a property of tests (namely, that these tests measure what they should measure)" (Borsboom et al., 2009, p. 138), then non-measurement based ascriptions of validity might, incorrectly,

Table 1
Validity as referring to interpretations of test scores in key APA texts.

Source	Quote
Test validity chapter in the *APA Handbook of Testing and Assessment in Psychology* (Sireci & Sukin, 2013, p. 61)	"[W]hen evaluating a test, the test itself is not what is evaluated or validated; rather, the interpretations or decisions derived from test scores are what must be validated. Therefore, the concept of validity is comprehensive and refers not only to test characteristics but also to the appropriateness of test use and to the accuracy of the inferences made on the basis of test scores."
Standards for Educational and Psychological Testing (AERA et al., 2014, p. 11)	"Statements about validity should refer to particular interpretations for specified uses. It is incorrect to use the unqualified phrase 'the validity of the test'"
Journal Article Reporting Standards for Quantitative Research in Psychology: The APA Publications and Communications Board Task Force Report (Appelbaum et al., 2018, p. 9)	"[T]he term *validity* should refer not to a test but to the proposed interpretations of test scores."

[5] In this paper, we frequently refer to the use of "tests" and "test scores" in our discussion of measurement procedures and measurement outcomes. However, our arguments are intended to cover the full range of psychological measurement procedures (e.g., performance-based, self-report, and physiological measures) and their outputs.

[6] As we discuss in Section 3, validity can also be ascribed to test scores under this conception of validity, whereas the concept of validity put forth by construct validity theorists ascribes validity to interpretations of test scores.

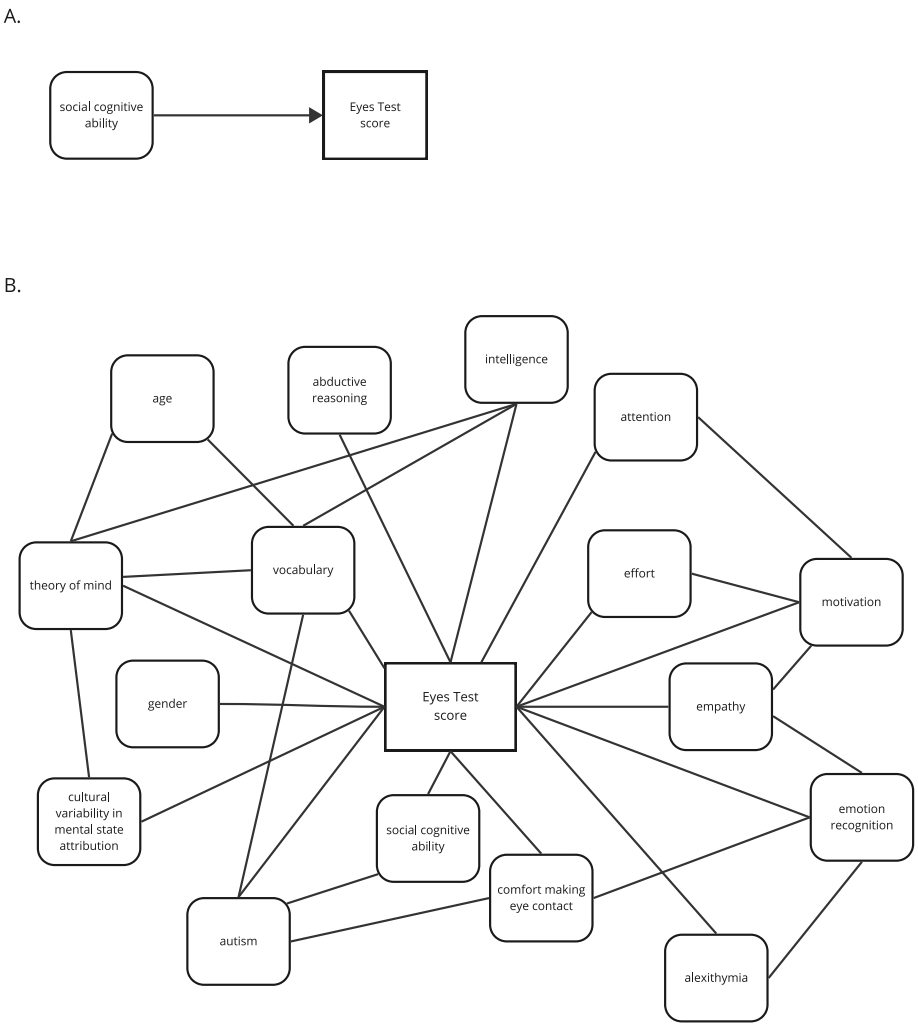


Fig. 2. Causal Complexity and Eyes Test Performance. (A.) The necessary causal relationship for the Eyes Test to be a valid measure of social cognitive ability. (B.) Illustrative examples of other factors that might be causally relevant for Eyes Test performance drawn from the literature surveyed by Higgins and colleagues (2024). We have not differentiated causal from non-causal associations in 2B (hence, the absence of arrowheads). Rather, we illustrate that a complex web of likely associations needs to be considered when evaluating psychological measurements as causal inferences.

be taken to indicate valid *measurement*.

A second benefit of a measurement-specific concept of validity is that it can facilitate the development of clear guidelines for measurement validation that are based specifically on the necessary and sufficient conditions for valid measurement. As things stand, it has been argued that "despite [validity's] centrality to the scientific enterprise, few, if any, clear standards regarding construct validation have been proposed" (Grimm & Widaman, 2023, p. 769). Moreover, it has been argued that the lack of guidelines is due, at least in part, to an overly broad definition of validity. For example, according to Grimm and Widaman (2023), one cause of the lack of validity standards is the multifaceted nature of validity when it includes different interpretations and uses of test scores. In a similar vein, Borsboom and colleagues (2004) argue that validity encompasses too many test-related considerations, with the result that the concept of validity provides limited practical guidance for researchers trying to validate measurements. And Cizek (2012) argues that the concept of validity "has suffered" because it refers to both the interpretation of the meaning of test scores and the intended use of test scores, which are, in his words, "important but incompatible dimensions" (p. 35).

As an illustration of how the inclusion of non-measurement-based interpretations and uses of test scores within the purview of validity may be inhibiting the development of clear guidelines for the validation of psychological measurements, we return to the distinction between measurement and prediction. As we have already discussed with the example from the *Standards* about a test of mathematics achievement being used to predict college-level success, a test's ability to measure an attribute and its ability to predict an outcome can vary independently. Even more problematically, there can be a trade-off between a test's ability to accurately measure a construct and its ability to accurately predict an outcome (Revelle, 2024; Smits et al., 2018). One reason for this is that the prediction of an outcome can be enhanced by combining multiple indicators (e.g., test questions) that each correlate strongly with the outcome of interest and only weakly with each other (e.g., because they measure different things). For example, to predict how well a student will perform on an exam, combining multiple different indicators that are associated with exam performance but not necessarily associated with each other (e.g., time spent studying, hours of sleep, and performance on previous assessment tasks) will likely increase the accuracy of the prediction versus having multiple indicators of a single factor (e.g., self-reported sleep duration, data from a consumer sleep tracking device, and electroencephalogram readings as indicators of hours of sleep).

Conversely, if multiple indicators are to be used to measure a single

attribute, the indicators should correlate with each other and the attribute being measured. For example, we would expect the three indicators of hours of sleep listed above to strongly correlate with each other, otherwise, it would be difficult to claim that they each accurately measure sleep duration. In the context of psychological tests, multiple items are often combined into a single score that is supposed to measure a specific psychological attribute. The extent to which these test items covary is referred to as internal consistency or internal reliability, and internal reliability is often considered a necessary condition for valid measurement when multiple items are summed into a single score (e.g., McNeish, 2024; Slaney, 2017). However, because internal reliability can be antagonistic to prediction, a guideline stipulating that internal reliability is necessary for validity would be inappropriate under a concept of validity that includes both prediction and measurement. This is problematic because it can create the impression that internal reliability is not important for valid measurement.

An example of internal reliability evidence being discounted in the context of measurement can be seen in a debate about the validity of Eyes Test scores. Higgins and colleagues have argued in several papers that Eyes Test scores are unsuitable as measurements of social cognitive ability based, in part, on evidence of unacceptably low levels of internal reliability and very weak correlations between test items (Higgins et al., 2024; Higgins et al., 2025; Higgins, Savalei, et al., 2023; 2025). Murphy and Hall (2024) have cautioned against this conclusion arguing that test scores with low levels of internal reliability "can [still] have strong value in *predicting* external variables" (p. 3; emphasis original). We maintain that this exchange illustrates how a concept of validity that encompasses both measurement and prediction can lead to a lack of clarity about the types of evidence required to validate the use of test scores: Murphy and Hall (2024) have discounted an important source of validity evidence in relation to measurement because it does not apply to prediction.

Based on the above considerations, we contend that psychological research would benefit from a measurement-specific concept of validity. This could be achieved by limiting the purview of validity to measurement. Alternatively, under a broad concept of validity that includes non-measurement related interpretations and uses of test scores, "measurement validity" could be a subtype of validity with its own validation guidelines. Having made our case for a measurement-specific concept of validity, in Section 3, we turn to the conditions necessary for valid measurement to obtain. In particular, we build on the position that a causal relationship from the attribute being measured to the measurement outcome is necessary for valid measurement (Borsboom et al., 2004; Eronen, 2025).

3. A causal condition for Valid measurement

Drawing on causal theories of measurement (e.g., Trout, 1999), Borsboom and colleagues' (2004) concept of validity is based on a causal relationship from the attribute being measured to the measurement outcome. While noting a dearth of explicit references to causality by leading validity theorists, they argue that earlier work on validity — including Loevinger's (1957) seminal paper on construct validity — have "hinted at" (p. 1062) a causal requirement for valid measurement. Moreover, they suggest that "it is likely that most researchers think of construct validity in causal terms, so that one could consider the proposed conception [of validity] to be a kind of underground interpretation of construct validity" (p. 1062). Focusing on content in the *Standards* specific to measurement (versus other interpretations of test scores), we likewise see evidence of an implicit causal interpretation of valid measurement. For example, the *Standards* state that "[i]f the test specification delineates the [psychological] processes to be assessed, then evidence is needed that the test items do, in fact, *tap* the intended processes" (AERA et al., 2014 p. 21, emphasis added). A natural interpretation of what it means for a test item to "tap" a psychological process is that performance on the test item is *caused* by the psychological process being tapped.[7] Moreover, this quote states that evidence for this (presumably causal) relationship should be provided. Another example from the *Standards* implies that a valid assessment of mathematics reasoning requires test item performance to be caused by mathematics reasoning:

For instance, if a test is intended to assess mathematics reasoning, it becomes important to determine whether test takers are, in fact, reasoning about the material given instead of following a standard algorithm applicable only to the specific items on the test (AERA et al., 2014, p. 15).

These examples point to a causal relationship from attributes to measurement outcomes being necessary for valid measurement according to the *Standards*, even though causality is not explicitly mentioned.[8] Whether explicit or underground, given that we define measurement in terms of quantifying the amount of a given attribute an entity possesses (see footnote [1]), we agree that causality is a necessary condition for valid measurement. If the amount of the attribute can change (by more than the level of precision of a given measurement procedure) without causing changes to the measurement outcome, how can the measurement procedure provide meaningful quantitative information about the attribute? We can *estimate* the level of an attribute based on non-causal sources of information, including measurements of attributes that are believed to share a common cause with the attribute of interest or that have been observed to consistently covary with the attribute of interest. However, for a test to *measure* an attribute, we maintain that changes in the attribute must cause changes in the measurement outcomes.

Building on the work of Borsboom and colleagues (2004), Eronen (2025) recently proposed a "*minimal causal condition for measurement*: *O* is a valid measure of *X only if* there is a causal relationship from *X* to *O*" (emphasis in original, p. 2220). While similar to the concept proposed by Borsboom and colleagues (2004), it differs in three important respects. First, Borsboom and colleagues' (2004) concept of validity requires a commitment to realism about measured attributes: "The attribute to which the psychologist refers must exist in reality; otherwise, the test cannot possibly be valid for measuring that attribute … Thus, measurement is considered to involve realism about the measured attribute" (p. 1063). By contrast, Eronen adopts a interventionist account of causal relevance (Pearl, 2009; Woodward, 2003), which is widely held to sidestep metaphysical issues concerning causality and the ontological status (i.e., existence) of the attribute being measured. In particular, Woodward (2015) presents his interventionist account "as a set of *methodological* proposals about explanation and causal inference, rather than as a set of theses about the *ontology* or *metaphysics* of causation" (p. 3577, emphasis in original).

The motivating idea, embraced by philosophers (e.g., Woodward, 2003) and scientists alike (e.g., Cook & Campbell, 1979, pp. 1–36), is that a causal relationship is one in which "manipulation of a cause will result in the manipulation of an effect" (Cook & Campbell, 1979, p. 36). The central notion for many interventionist accounts is that of a "direct

[7] This is another example of how the specific requirements for valid measurement can get overlooked or disregarded when validity also includes non-measurement-based interpretations of test scores.

[8] We suggest that the absence of causal language in the *Standards* is the result of a widespread "taboo" against explicit causal inferences in nonexperimental psychological research (Grosz et al., 2020; see also Borsboom et al., 2004; Eronen, 2025). Indeed, it has been convincingly argued that this taboo frequently prompts a shift from explicit to implicit expressions of causal inferences rather than a shift away from the causal inferences themselves (Grosz et al., 2020).

cause". According to Woodward (also see Pearl, 2009), "a necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by interventions" (Woodward, 2003, p. 55). This conception is intended to capture the essence of a carefully designed, randomised experiment because any statistically significant differences observed in the outcome variable are likely to be the result of the experimental intervention, and not changes in various other confounding variables that might be correlated with, but lie on indirect causal paths to, the outcome variable. For the purpose of informing psychological research practices, we think it is both acceptable and preferable to adopt an interventionist account of causal relevance that captures scientific norms and does not require committing to particular metaphysical and ontological theories of causation or causes (Eronen, 2025; Woodward, 2015).

The second difference between the two concepts of validity is that, unlike Borsboom and colleagues (2004), Eronen (2025) argues that a causal relationship from an attribute to a measurement outcome is *not sufficient* for valid measurement. This is an important distinction to make in relation to the measurement of psychological attributes due to causal complexity. As an illustration of how causal complexity makes a necessary and sufficient causal condition for valid measurement undesirable, consider factors that might influence performance on the Eyes Test. For the Eyes Test to satisfy a necessary and sufficient causal condition for valid measurement of social cognitive ability, there must be a causal relationship from social cognitive ability to Eyes Test scores (as represented in Fig. 2A) such that a hypothetical intervention on social cognitive ability would cause changes to Eyes Test scores. Of course, Fig. 2A is a vast oversimplification. In practice, Eyes Test scores are likely directly and indirectly causally influenced by numerous factors other than social cognitive ability (see Fig. 2B), which *may* include differences in vocabulary (Olderbak et al., 2015); differences in level of comfort viewing eye stimuli (Higgins, Ross, et al., 2023); differences in abductive reasoning ability (Higgins et al., 2025); culturally variable norms and practices around mental state attribution (Luhrmann et al., 2011); levels of attention, motivation, and effort during task completion (Stosic et al., 2024); and so on. This makes Fig. 2B a more realistic representation of the potential factors associated with, and possibly causally related to, Eyes Test scores.

Like Eyes Test scores, output from other psychological measurement procedures will almost certainly sit at the centre of complex webs of causal pathways and associations. Consequently, few (if any) psychological measurements are only, or even predominantly, caused by the attribute they are intended to measure. As such, it would be undesirable to embrace a concept of validity that classifies test scores as valid measurements of *every* factor that exerts a causal influence on them, regardless of the strength of the causal influence (Larroulet Philippi, 2021).[9] Yet, this would have to be the case if a causal relationship from a given attribute to test scores were *sufficient* for valid measurement.

Eronen's (2025) move to make a causal relationship from attributes to measurement outcomes necessary but not sufficient for valid measurement addresses the issues associated with classifying tests as valid measures of all constructs that exert a causal influence on test scores. However, it introduces its own complications. Consider Borsboom and colleagues' (2004) first necessary condition for valid measurement (i.e., that the measured attribute must exist). There is undoubtedly merit to the argument that, to exert a causal influence, a psychological attribute must exist (see Hood, 2009). However, it does not follow from this that establishing the existence of a psychological attribute (rather than assuming its existence) must be a focus of measurement validation. Indeed, as argued by Eronen (2025), this condition can be sidestepped by adopting the interventionist account of causal relevance. Similarly, if causality is necessary, but not sufficient, for valid measurement, then it is not clear that explicating causal relationships from attributes to measurement outcomes — rather than assuming the necessary causal relationship exists — should be the focus of measurement validation. Moreover, it leaves us with the question of what *is* sufficient for valid measurement.[10]

We propose that there is a middle ground that addresses both these issues, whereby a causal relationship from the attribute being measured to measurement outcomes is *in principle* sufficient for valid measurement. In particular, a causal relationship is sufficient for valid measurement if it is strong enough that changes to the attribute above the level of precision of the measurement procedure would cause changes to the measurement outcome that are proportional to the changes in the attribute.[11] Adopting a necessary, and in principle sufficient, causal condition for valid measurement secures causality's position at the centre of the measurement validation process.

Critically, if valid measurement requires a (sufficiently strong) causal relationship from attributes to measurement outcomes, then a causal inference is made every time someone claims to have measured something, whether explicitly recognised or not. If these causal inferences are made implicitly, they cannot be properly evaluated, which can result in flawed or unjustified causal inferences going undetected (Bulbulia, 2024; Rohrer, 2018). In the context of empirical research, this could translate to invalid measurements being uncritically accepted as valid. By contrast, making the development and refinement of causal models for measurement procedures — and the accrual of evidence in support of those models — central to the validation process can provide a structured approach through which researchers can critically assess psychological measurements as causal inferences. It also highlights the need to consider how measurement procedures can instantiate causal relationships from attributes to measurement outcomes when developing psychological measurement tools (Borsboom, 2005). Of course, developing and evaluating causal models for psychological measurement procedures will be extremely challenging, to say the least (Eronen, 2020, 2025). We return to this issue in Section 5.

A third important difference between Eronen's (2025) and Borsboom and colleagues' (2004) concepts of validity is that Eronen ascribes validity to test scores while Borsboom and colleagues ascribe validity to tests.[12] Unlike the debate about ascribing validity to tests versus *interpretations* of test scores, we are not aware of any discussions in the literature about a distinction between the validity of tests and the validity of test scores; nor does Eronen (2025) draw this distinction between his and Borsboom and colleagues' concepts of validity. In fact, Eronen (2025) summarises Borsboom and colleagues' (2004) concept of validity in the following way: "a *test or measurement* is valid if and only if it measures what it is intended to measure" (p. 2219, emphasis added). This subsumption of tests and measurements under a single concept of validity is consistent with our impression (without having systematically reviewed the literature) that the validity of tests and the validity of test scores are often lumped together and treated interchangeably. In Section 4, we argue that it is important to make an explicit distinction between the "validity" of tests and the "validity" of test scores, and we describe four benefits of this distinction in the context of psychological research.

[9] Variations in causal strength can be conceptualised as differences in the probability that a change to one variable will result in a change to another variable (Fitelson & Hitchcock, 2011; Pearl, 2009).

[10] We thank a reviewer for pointing out this issue.

[11] Or, in some cases, the intervention might cause changes to the probability distribution of the attribute.

[12] Taken in isolation, Eronen's minimal causal condition also appears to ascribe validity to tests given the use of the phrase "*O* is a valid *measure* of *X*" (p. 2220, emphasis added) rather than *O is* a valid *measurement* of *X*. However, Eronen (2025) specifies that *O* refers to test scores: "*O* can refer to any outcome of a measurement procedure, such as instrument readings or responses to items in a questionnaire" (p. 2220).

4. Validity potential versus realised validity

It is important to distinguish between what we refer to as the validity potential of a measurement procedure *in abstracto* (e.g., the Eyes Test itself) and the realised validity of the output from each instantiation of that measurement procedure (e.g., Eyes Test scores collected in a particular study). As we define it, validity potential is the probability that a measurement procedure will produce valid measurements of the attribute it is intended to measure, such that tests with higher validity potential have a greater probability of producing valid measurements.[13] By contrast, realised validity concerns specific measurement outcomes and whether they satisfy the criteria for valid measurement. In other words, a measurement with realised validity is a valid measurement. Framed in terms of our necessary and in principle sufficient condition for valid measurement:

> *O* is a valid measurement of *X* if and only if changes to *X* above the level of precision of the measurement procedure cause changes to *O* that are proportional to the changes in *X*.

To illustrate this distinction, consider thermometers. There are several types of thermometers that have very high levels of validity potential, including mercury, infrared, and digital thermometers. This means that a given temperature reading from these types of thermometers is highly likely to be a valid measurement of temperature (i.e., the measurement will have realised validity). Nonetheless, they do not always produce valid measurements. For example, if the mercury in a mercury thermometer freezes, it will not produce valid measurements, or if an infrared thermometer is held too far from a person's forehead, the temperature reading will be less accurate. Thus, the validity potential of a thermometer and the realised validity of a given output from that thermometer are distinct.

The high validity potential of thermometers reflects centuries of research and scientific advances that have culminated in: (1) a well-developed theory of temperature, (2) an understanding about how changes in temperature can reliably produce changes in temperature readings for each type of thermometer, and (3) a large body of empirical evidence documenting the contexts in which thermometers reliably produce valid temperature readings and those in which they do not (i.e., the boundary conditions of reliable operation; Chang, 2004). Critically, our understanding of temperature, how thermometers work, and their boundary conditions allows us to estimate their validity potential.

Generalising from thermometry, we propose three key factors that determine the validity potential of a measurement procedure, which derive from both theoretical and empirical sources. The first factor is the current state of knowledge about the attribute being measured. The better we understand an attribute, the better able we will be to design and evaluate measurement procedures for that attribute. The second factor, which relies heavily on our understanding of the relevant attribute, is our ability to explain *how* a measurement procedure can instantiate a causal relationship from the attribute to the measurement outcomes. This can take the form of a causal model of the measurement procedure (see Section 5). The third factor is empirical evidence demonstrating the range of contexts in which a given measurement procedure can reliably produce valid measurements.[14]

In the context of psychological measurement, we see at least four benefits to making the distinction between validity potential and realised validity explicit. The first benefit is that validity potential makes it clear that theoretical and conceptual work, which have been underemphasised in the psychological sciences (Bringmann et al., 2022; Muthukrishna & Henrich, 2019), are critical for valid measurement and, thus, critical for empirical research. As it stands, the validity potential of many psychological measurements is likely very low because we do not have well developed theories of psychological attributes (Muthukrishna & Henrich, 2019). Rather, there is a widespread, and underappreciated, lack of conceptual clarity in the psychological sciences (Bringmann et al., 2022), and there are few instances of attempts being made to explain how psychological tests might instantiate a causal relationship from attributes to measurement outcomes. Increasing the validity potential of psychological measurement procedures will require a substantial commitment to clarifying psychological concepts and explicating how measurement procedures can instantiate a causal relationship from attributes to measurement outcomes.[15]

A second benefit of an explicit distinction between validity potential and realised validity is that it reinforces the importance of sample-specific validity evidence. This dovetails nicely with existing recommendations from the APA regarding the provision of sample-specific validity evidence in every study (e.g., Appelbaum et al., 2018), which currently are largely overlooked (e.g., Flake & Fried, 2020). By contrast, the poor validity evidence reporting practices observed across many areas of psychological research (see Flake & Fried, 2020) mean that there is currently little evidence that psychological tests have reliably produced measurements with realised validity. This limits the validity potential of psychological measurement procedures since, as we define it, evidence of realised validity is a key contributing factor to validity potential.

A third benefit of an explicit distinction between validity potential and realised validity is that they map onto different stages of the research process. Validity potential is critical during the design phase of a study when measurement procedures are selected. At this stage, researchers should consider whether the validity potential of candidate measures make it reasonably likely that the measurements collected within their study will be valid, paying particular attention to any relevant known or probable confounding factors. Careful consideration of validity potential prior to data collection can help avoid wasting resources on studies where weak validity potential makes it unlikely that valid measurements will be obtained. For example, a researcher might rule out the use of the Eyes Test because the weakness of the validity evidence summarised by Higgins and colleagues (Higgins et al., 2024; Higgins et al., 2025) and concerns they raised about the design of the test indicate that the Eyes Test generally has very low levels of validity potential. By contrast, realised validity is assessed after the data are collected to determine whether the specific measurements collected in the study are valid. Realised validity is critical for making inferences based on measurements and for evaluating the credibility of those inferences.

A fourth benefit of an explicit distinction between validity potential and realised validity is that validity potential can accommodate the intuition that tests themselves can be valid (Borsboom et al., 2004; Hood, 2012) while avoiding the problematic inference that tests have context-independent or "unconditional validity" (Larroulet Philippi, 2021, p. 164; see also Newton, 2012). As illustrated by the example of the frozen mercury thermometer, even measurement procedures with very high levels of validity potential are not unconditionally valid. Nonetheless, if validity is ascribed to tests in the abstract, it leaves

[13] Here we focus on the use of tests for measurement. However, this distinction is also relevant for other uses of psychological tests including prediction and diagnosis. If this distinction were to be adopted across test uses, it might clarify things to refer to specific subtypes of validity potential such as "measurement potential", "predictive potential", and "diagnostic potential".

[14] We thank a reviewer for pointing out that the distinction between validity potential and realised validity could be captured in Bayesian terms. We agree that this would be an interesting direction to pursue.

[15] While beyond the scope of the current paper, our proposed distinction between validity potential and realised validity is highly compatible with Chang's (2004) epistemic iteration. In particular, validity evidence can be used iteratively to inform modifications to measurement procedures and theories of psychological constructs, which can lead to improvements in the theories, measurement procedures, and measurement outcomes.

considerable scope for researchers to uncritically assume that measurements from a test that has been labelled "valid" are valid. In fact, this would seem a natural interpretation if validity is conceptualised as "a property of tests (namely, that these tests measure what they should measure)" (Borsboom, 2005, p. 138). Viewed from the perspective of unconditional validity, it seems likely that the low levels of reporting of sample-specific validity evidence for the Eyes Test (Higgins et al., 2024) is due, at least in part, to a widespread perception that the Eyes Test is a valid measure, and, therefore, produces valid measurements. However, despite the problems associated with unconditional validity, measurement procedures must have at least *some* level of context independent validity or else we could never generalise research findings. Validity potential can capture both context-dependent and context-independent features of measurement procedures, with both contributing to the validity potential of the measurement procedure. This works because the three factors that contribute to validity potential have varying levels of context-dependency. The theory of the attribute should be relatively context independent. The proposed causal relationship from the attribute to the measurement outcomes should also have a relatively high level of context-independence. For example, the explanation for how a mercury thermometer causes changes in the temperature reading is constant across contexts, and can predict contexts in which confounding contextual factors will reduce the validity potential of the thermometer. The empirical evidence based on the realised validity of actual outcomes from the measurement procedure adds context-specific information to validity potential. In Section 5, we turn to the question of how to validate psychological measurements as causal inference.

5. Validating psychological measurements as causal inferences

It is extremely challenging to make well-founded causal inferences (Bulbulia, 2024; Eronen, 2025; Pearl, 2009), and thus, the research community should accept that the project to validate psychological measurements as causal inferences will also be extremely challenging. In fact, despite proposing a minimal causal condition for valid measurement, Eronen (2025) argues that most current psychological measurement cannot meet this requirement, due in large part to the causal complexity of human psychology and the fact that most psychological constructs are ill-defined. Moreover, Eronen (2020, 2025) argues that attempts to identify psychological causes face serious obstacles, even within an interventionist framework, because, in addition to the challenges posed by causal complexity and ill-defined constructs, psychological interventions are "fat-handed." This means that any given intervention on a psychological cause is likely to impact multiple factors simultaneously. He further argues that it is unlikely that confounding psychological factors can be held fixed. To address these issues while avoiding problems associated with declaring most psychological measurements invalid, Eronen (2025) proposes a distinction between "hard" and "soft" measurement. "Hard" measurement is causal, whereas "soft" measurement does not meet the minimal causal condition but is nonetheless useful. In addition to allowing for psychological measurements to be valid in a weaker sense, he argues that distinguishing between hard and soft measurement can draw attention to the need to better understand what psychological measurement *is* (rather than rejecting it outright for not meeting the criteria of "genuine" measurement).

As an example of weak measurement, Eronen (2025) suggests that intelligence tests could have "a degree of validity for predicting academic achievement" (p. 2226) based on correlations between intelligence test scores and measures of academic achievement, without requiring a direct causal relationship from intelligence to intelligence test scores.[16] However, in line with our discussion in Section 2 about the differences between measurement and prediction, we argue that the reason that causality is not necessary in this case is that this use of the test does not require that the test *measures* intelligence. Under our account, what matters is that the correlation between the test scores and academic achievement is reliable enough for the scores to accurately *predict* academic achievement, irrespective of what the test measures. Alternatively, if the test scores were to be used to make claims about the relationship between intelligence and academic achievement, then the test would need to measure intelligence, and the causal condition would apply. Consequently, while we agree with Eronen (2025) that it is important to gain a better understanding of what "psychological measurement" is (and is not) and that psychological data can have uses other than measurement, we disagree that the introduction of "soft measurement" is the best way to address this issue. Instead, we contend that it is more useful to make a distinction between uses of test scores that require the assumption of measurement and those that do not – and to label and evaluate those uses accordingly. Importantly, if a causal relationship from attributes to measurement outcomes is necessary for valid measurement, yet most psychological measurements cannot satisfy this condition, the psychological research community has a serious problem that it needs to tackle head on. In line with Eronen's concerns, as part of a process of critical reflection, the research community should be open to the possibility that some psychological attributes cannot be measured, either because we do not currently possess appropriate technology or sufficient theoretical understanding, or because the nature of some psychological attributes makes them unmeasurable (Michell, 2021). Crucially, not being measurable "does not mean that such attributes cannot be investigated scientifically, only that they cannot be measured" (Michell, 2021, p. 16). Nonetheless, we think it would be premature to conclude that, if we accept a necessary causal condition for valid measurement, psychological measurement is in principle impossible.

As a starting point for attempting to validate psychological measurements as causal inferences, we suggest turning to the rich existing and emerging literature on causality and causal inferences in empirical research (e.g., Bulbulia, 2024; Pearl, 2009; Rohrer, 2024). Returning to the example of the Eyes Test as a putative measure of social cognitive ability, to satisfy a causal condition for valid measurement (i.e., changes in social cognitive ability *cause* changes in Eyes Test scores) within an interventionist framework, there must, in principle, be a manipulation that changes levels of social cognitive ability, and the change in social cognitive ability must result in a change in Eyes Test scores when all other relevant factors are held constant. Thus, an important part of the validation process should be identifying relevant causal factors, assessing their roles relative to the causal relationship of interest, and from there, considering how best to assess whether the necessary causal condition is satisfied. Directed acyclic graphs (DAGs[17]; Bulbulia, 2024; Pearl, 2009; Rohrer, 2024; Woodward, 2021) are one tool that could help researchers develop and test plausible, yet tractable, causal models for measurement procedures within the context of the complex webs of associations in which psychological measurements are likely to sit (e.g., Fig. 2B).[18] DAGs can be useful when assessing validity evidence because they make the assumptions of the causal model for the measurement procedure explicit (Rohrer, 2018).

Among other things, DAGs can be used to visualise three key roles that variables can play relative to a causal relationship between two

[16] This is equivalent to the example, discussed in Section 2, from the *Standards* of using a mathematics achievement test to predict success at college-level work.

[17] DAGs constitute simplified models of reality in which variables are represented as nodes and causal relationships as arrows between nodes; they are acyclic because they do not contain causal cycles (i.e., there will never be a causal chain from a variable back to itself).

[18] Although beyond the scope of this paper to address, we acknowledge that there is another epistemic problem lurking here pertaining to variable choice: how to decide which variables should be included in a given causal model from the vast number of candidate variables. See Woodward (2016) for further discussion.

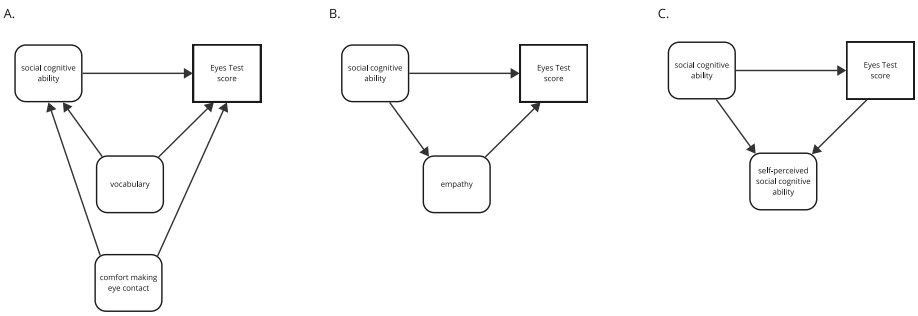


Fig. 3. Modelling Potential Causal Roles of Variables Relative to Eyes Test Score. (A) DAG representing vocabulary and comfort making eye contact variables as "confounders" for the relationship between social cognitive ability and Eyes Test scores variables. (B) DAG representing the empathy variable as a "mediator" for the relationship between social cognitive ability to Eyes Test scores variables. (C) DAG representing the self-perceived social cognitive ability variable as a "collider" for the relationship between social cognitive ability and Eyes Test scores variables.

variables of interest (Pearl, 2009; Rohrer, 2024). This is important because a variable's role relative to the causal relationship of interest determines whether it is desirable to control for changes in that variable. First, a variable is a "confounder" if it results in a non-causal association between two variables of interest, making it difficult to assess the strength of the causal relationship of interest. For example, in Fig. 3A, we model "vocabulary" and "comfort making eye contact" as confounders relative to the hypothesised causal relationship from social cognitive ability to Eyes Test scores. Controlling for confounders may help avoid spurious correlations. Second, a variable is a "mediator" if it transmits the causal effect between two variables of interest. In Fig. 3B, we model "empathy" as a mediator relative to the causal relationship from social cognitive ability to Eyes Test scores. Controlling for mediators can reduce the overall association between the variables of interest, leading to an underestimation of the strength of the causal relationship. Third, a variable is a "collider" if it is a common effect of the two variables of interest. In Fig. 3C, we model "self-perceived social cognitive ability" as a collider relative to the relationship from social cognitive ability to Eyes Test scores, under the assumption that both levels of social cognitive ability and performance on a test that is intended to measure social cognitive ability can cause changes to a person's perception of their own social cognitive ability. Controlling for colliders can create spurious associations between variables of interest. Thus, in addition to identifying relevant causal variables, it is critical to consider the role that each variable might play relative to the causal relationship of interest. Breaking down complex webs of associations (e.g., Fig. 2B) into DAGs representing these roles can help us determine which variables we need to attempt to control for when validating psychological measurements.

Most psychological attributes are likely difficult, if not impossible, to manipulate while holding *all* other causally relevant factors fixed (Eronen, 2025), which presents a substantial obstacle to evaluating proposed causal models of measurement procedures. Researchers could attempt to intervene on psychological attributes using methods that aim to induce temporary changes to psychological attributes, such as transcranial magnetic stimulation (TMS) and transcranial direct current stimulation (tDCS). For example, it has been argued that these techniques applied over different areas of the brain can temporarily increase or decrease emotion recognition ability (Andò et al., 2021). However, Eronen's (2020) concerns about the fat-handedness of interventions would be highly relevant here given that these types of brain stimulation are unlikely to be precise enough to avoid impacting brain areas associated with psychological attributes other than the target attribute.

Sources of validity evidence that do not employ manipulations can also contribute to the evaluation of causal models of measurement procedures. These sources of validity evidence will be imperfect. Nonetheless, in combination, imperfect sources of evidence can potentially converge toward what Wimsatt (2007) refers to as "robustness" (see also Bringmann & Eronen, 2016; Eronen, 2015). Alternatively, when evidence does not converge, it can be used to refine or reject the causal model or measurement procedures. This use of imperfect, converging sources of evidence to support the validity of psychological measurements is already a key recommendation of current validation guidelines, but without the explicit focus on causal relationships. Here we briefly consider how three widely used sources of validity evidence described in the *Standards* and seminal works on construct validation (see Flake et al., 2017 for a summary) could be interpreted in the context of assessing causal models of measurement procedures, while highlighting key limitations with each type of evidence.

The first source of validity evidence is factor analysis, which is used to evaluate the factor structure of test scores based on patterns in the covariances (e.g., correlations) between test items. The aim of factor analysis is to identify patterns that are consistent with the existence of one or more latent (i.e., unobservable) factors that influence (or cause) performance on test items. In the case of the Eyes Test, all items are intended to measure a single psychological attribute, thus, we would expect the pattern of covariance between items to be consistent with a single-factor model in which a single latent factor influences performance on all test items. By contrast, if a test is designed to measure multiple attributes, we would expect the pattern of covariance between items to be consistent with the existence of multiple latent factors. Key limitations of factor analysis are that it cannot tell you *what* a test measures (Watts et al., 2023) and multiple factor models can potentially explain the same dataset similarly well (van Bork et al., 2017).[19]

A second source of validity evidence is known groups validity evidence. In known groups validity evidence, the average scores of two groups that are "known" to differ on the specific attribute that a test is intended to measure are compared to see whether the average scores of each group are consistent with these "known" differences. For example, alexithymia is a subclinical condition characterised by difficulty recognising one's own emotional states, which is associated with poorer ability to recognise emotions in other people (Oakely et al., 2016). In the context of assessing the Eyes Test as a measure of emotion recognition ability, if differences in emotion recognition ability cause differences in Eyes Test scores, then we would predict that the average Eyes Test score would be lower for a group of people with alexithymia than for a group of people without alexithymia. A key limitation of known groups validity evidence is that it relies on the assumptions that the two groups of people actually differ in terms of the particular psychological attribute of interest *and* that differences in average test scores between the two groups are caused by this difference in the target attribute. However, in practice, performance differences might be caused by construct-irrelevant factors (AERA et al., 2014).

[19] The insensitivity of factor analysis results to specific sources of variance was strikingly demonstrated by Maul (2017), who showed that even responses to Likert scales comprising nonsense items can have an excellent factor structure.

A third source of validity evidence is convergent validity evidence. In the case of convergent validity evidence, strong correlations with another test that is intended to measure the same construct support the validity potential of a measure. In the context of a causal relationship from attributes to measurement outcomes, measurements from two measures of the same construct should be highly correlated because changes to the target construct should cause changes to the measurement outcomes from both measures. Limitations of convergent validity evidence include its reliance on the assumption that the alternative test produces valid measurements of the target attribute and the assumption that correlations in performance on the two tests occur *because* both tests measure the attribute of interest (Kellen et al., 2021).

While far from comprehensive, in this section we have illustrated how the psychological research community might begin to approach the challenge of validating psychological measurements as causal inferences using existing methods and tools. Given the challenges posed by causal complexity, a lack of conceptual clarity, and the fat handedness of psychological interventions (Bringmann et al., 2022; Eronen, 2020, 2025), we anticipate that the results will be disheartening and that assessments of psychological measurements as causal inferences will indicate that many existing psychological measurements are not valid and many psychological measurement procedures have very low validity potential. However, if this proves to be the case, the scientific merit of psychological research will benefit in the long run if we take the necessary steps to develop better measures and/or explore alternative methods for studying psychological phenomena.

6. Conclusion

In this paper, we proposed a concept of validity that we argued has greater practical utility than the concept of validity underlying influential best practice guidelines. Drawing on the widely used but inadequately validated Eyes Test as an example of validation failure, we described how three key features of our proposed concept of validity can encourage better validation practices. The first key feature is that our concept of validity is explicitly restricted to measurement, which we argued can concomitantly make the requirements for valid measurement clearer and make validity claims easier to interpret. The second key feature is a causal condition that is both necessary and, in principle, sufficient for valid measurement. Although we acknowledged that causal complexity poses a significant barrier to achieving valid psychological measurement given a necessary causal condition, we highlighted how the causal inference literature can serve as a starting point for evaluating psychological measurements as causal inferences. The third key feature is that our concept of validity makes an explicit distinction between validity potential, which applies to measurement procedures *in abstracto*, and realised validity, which applies to the output from specific instantiations of a measurement procedure. In closing, we encourage the psychological research community to prioritise the vital, yet extremely challenging tasks of developing measurement procedures with high levels of validity potential and routinely assessing the realised validity of measurement outcomes.

CRediT authorship contribution statement

Wendy C. Higgins: Writing – review & editing, Writing – original draft, Conceptualization. **David M. Kaplan:** Writing – review & editing, Conceptualization. **Alexander J. Gillett:** Writing – review & editing, Conceptualization. **John Sutton:** Writing – review & editing, Conceptualization. **Robert M. Ross:** Writing – review & editing, Conceptualization.

Declarations of interest

WCH was supported by an Australian Government Research Training Program (RTP) Scholarship, a Macquarie University Research Excellence Scholarship, and an Australian Research Council (ARC) Future Fellowship Grant (grant ID: FT210100652). AJG and JS were supported by the John Templeton Foundation (grant ID: 61924). RMR was supported by the John Templeton Foundation (grant ID: 62631).

Acknowledgements

We would like to thank Richard Menary, Hoda Mostafavi, Diana Tan, Friederike Charlotte Hechler, and the Theory and Methods in Biosciences Lab based at Macquarie University for their helpful feedback on previous versions of this manuscript.

Data availability

No data was used for the research described in the article.

References

Alexandrova, A., & Haybron, D. M. (2016). Is construct validation valid? *Philosophy of Science, 83*(5), 1098–1109. https://doi.org/10.1086/687941

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

Andò, A., Vasilotta, M. L., & Zennaro, A. (2021). The modulation of emotional awareness using non-invasive brain stimulation techniques: A literature review on TMS and tDCS. *Journal of Cognitive Psychology, 33*(8), 993–1010. https://doi.org/10.1080/20445911.2021.1954013

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42*(2), 241–251. https://doi.org/10.1017/S0021963001006643

Betz, N., Hoemann, K., & Barrett, L. F. (2019). Words are a context for mental inference. *Emotion, 19*(8), 1463–1477. https://doi.org/10.1037/emo0000510

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 135–170). Information Age Publishing.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to basics: The importance of conceptual clarification in psychological science. *Current Directions in Psychological Science, 31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology, 26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Bulbulia, J. A. (2024). Methods in causal inference. Part 1: Causal diagrams and confounding. *Evolutionary Human Sciences, 6*, 1–39. https://doi.org/10.1017/ehs.2024.35

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press. https://doi.org/10.1093/0195171276.001.0001

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43. https://doi.org/10.1037/a0026975

Cizek, G. J.. Validity an integrated approach to test score meaning and use. Routledge. https://doi.org/10.4324/9780429291661.

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment, 31*(12), 1412–1427. https://doi.org/10.1037/pas0000626.

Cook, T. D., & Campbell, D. T. (1979). Causal inference and the language of experimentation. *Quasi-experimentation: Design & analysis issues for field settings*.

Eronen, M. I. (2015). Robustness and reality. *Synthese, 192*(12), 3961–3977. https://doi.org/10.1007/s11229-015-0801-6

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology, 59*, Article 100785. https://doi.org/10.1016/j.newideapsych.2020.100785

Eronen, M. I. (2025). Causal complexity and psychological measurement. *Philosophical Psychology*, 2217–2232. https://doi.org/10.1080/09515089.2023.2300693

Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 600–627). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199574131.003.0029.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science, 3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science, 8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology, 1*(6), 358–368. https://doi.org/10.1038/s44159-022-00050-2

Grimm, K. J., & Widaman, K. F. (2023). Construct validity. In H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology: Foundations, planning, measures, and psychometrics* (2nd ed., pp. 769–791). American Psychological Association. https://doi.org/10.1037/0000318-035.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science, 15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2024). Construct validity evidence reporting practices for the reading the mind in the Eyes Test: A systematic scoping review. *Clinical Psychology Review, 108*. https://doi.org/10.1016/j.cpr.2023.102378

Higgins, W. C., Kaplan, D. M., Deschrijver, E., & Ross, R. M. (2025). Why most research based on the reading the mind in the Eyes Test is unsubstantiated and uninterpretable: A response to Murphy and Hall (2024). *Clinical Psychology Review, 115*. https://doi.org/10.1016/j.cpr.2024.102530

Higgins, W. C., Ross, R. M., Langdon, R., & Polito, V. (2023). The "Reading the Mind in the Eyes" Test shows poor psychometric properties in a large, demographically representative U.S. sample. *Assessment, 30*(6), 1777–1789. https://doi.org/10.1177/10731911221124342

Higgins, W. C., Savalei, V., Polito, V., & Ross, R. M. (2023). Validation of the reading the mind in the Eyes Test requires an interpretable factor model. *Proceedings of the National Academy of Sciences, 120*(52), Article e2303706120. https://doi.org/10.1073/pnas.2303706120

Higgins, W. C., Savalei, V., Polito, V., & Ross, R. M. (2025). Reading the mind in the Eyes Test scores demonstrate poor structural properties in nine large non-clinical samples. *Assessment*. https://doi.org/10.1177/10731911251328604

Hood, S. B. (2009). Validity in psychological testing and scientific realism. *Theory & Psychology, 19*(4), 451–473. https://doi.org/10.1177/0959354309336320

Hood, S. B. (2012). In defense of an instrument-based approach to validity. *Measurement: Interdisciplinary Research & Perspective, 10*(1–2), 63–65. https://doi.org/10.1080/15366367.2012.681976

Johnston, L., Miles, L., & McKinlay, A. (2008). A critical review of the Eyes Test as a measure of social-cognitive impairment. *Australian Journal of Psychology, 60*(3), 135–141. https://doi.org/10.1080/00049530701449521

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science, 16*(4), 767–778. https://doi.org/10.1177/1745691620974771

Kelley, T. L. (1927). *Interpretation of educational measurements*. World Book Company.

Larroulet Philippi, C. (2021). Valid for what? On the very idea of unconditional validity. *Philosophy of the Social Sciences, 51*(2), 151–175. https://doi.org/10.1177/0048393120971169

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Luhrmann, T., Astuti, R., & Robbins, J. (2011). Toward an anthropological theory of mind. *Suomen Antropologi: Journal of the Finnish Anthropological Society, 36*(4), 5–69.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspective, 15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

McNeish, D. (2024). Practical implications of sum scores being psychometrics' greatest accomplishment. *Psychometrika, 89*(4), 1148–1169. https://doi.org/10.1007/s11336-024-09988-z

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and Macmillan.

Michell, J. (2021). Representational measurement theory: Is its number up? *Theory & Psychology, 31*(1), 3–23. https://doi.org/10.1177/0959354320930817

Murphy, B. A., & Hall, J. A. (2024). How a strong measurement validity review can go astray: A look at Higgins et al. (2024) and recommendations for future measurement-focused reviews. *Clinical Psychology Review, 114*. https://doi.org/10.1016/j.cpr.2024.102506

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour, 3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

National Institute of Mental Health. (2016). Behavioral assessment methods for RDoC Constructs. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/constructs/understanding-mental-states.shtml.

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research & Perspective, 10*(1–2), 1–29. https://doi.org/10.1080/15366367.2012.669666

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Olderbak, S., Wilhelm, O., Olaru, G., Geiger, M., Brenneman, M. W., & Roberts, R. D. (2015). A psychometric analysis of the reading the mind in the Eyes Test: Toward a brief form for research and applied settings. *Frontiers in Psychology, 6*, 1503. https://doi.org/10.3389/fpsyg.2015.01503

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the Easter Bunny. *Personality and Individual Differences, 221*, Article 112552. https://doi.org/10.1016/j.paid.2024.112552

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rohrer, J. M. (2024). Causal inference for psychologists who think that causal inference is not for them. *Social and Personality Psychology Compass, 18*(3), Article e12948. https://doi.org/10.1111/spc3.12948

Schimmack, U. (2021). The validation crisis in psychology. *Meta-Psychology, 5*. https://doi.org/10.15626/MP.2019.1645

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton: Mifflin and Company.

Silverman, C. (2022). How to read 'Reading the Mind in the Eyes'. *Notes and Records of the Royal Society of London, 76*(4), 683–697. https://doi.org/10.1098/rsnr.2021.0058

Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger (Ed.), *1. Test theory and testing assessment in industrial and organizational psychology* (pp. 61–84). Washington, DC, US: American Psychological Association. APA handbook of testing and assessment in psychology.

Slaney, K. (2017). *Validating psychological constructs: Historical, philosophical, and practical dimensions*. Palgrave Macmillan.

Smits, N., van der Ark, L. A., & Conijn, J. M. (2018). Measurement versus prediction in the construction of patient-reported outcome questionnaires: Can we have our cake and eat it? *Quality of Life Research, 27*(7), 1673–1682. https://doi.org/10.1007/s11136-017-1720-4

Stosic, M. D., Murphy, B. A., Duong, F., Fultz, A. A., Harvey, S. E., & Bernieri, F. (2024). Careless responding: Why many findings are spurious or spuriously inflated. *Advances in Methods and Practices in Psychological Science, 7*(1). https://doi.org/10.1177/25152459241231581

Strohmaier, A. R., Reinhold, F., Hofer, S., Berkowitz, M., Vogel-Heuser, B., & Reiss, K. (2022). Different complex word problems require different combinations of cognitive skills. *Educational Studies in Mathematics, 109*(1), 89–114. https://doi.org/10.1007/s10649-021-10079-4

Tal, E. (2020). Measurement in science. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Fall 2020 ed. https://plato.stanford.edu/archives/fall2020/entries/measurement-science/

Trout, J. D. (1999). Measurement. In W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 265–276). Oxford, England: Blackwell.

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory & Psychology, 27*(6), 759–773. https://doi.org/10.1177/0959354317737185

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science: A Journal of the American Psychological Society, 31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Watts, A. L., Greene, A. L., Ringwald, W., Forbes, M. K., Brandes, C. M., Levin-Aspenson, H. F., & Delawalla, C. (2023). Factor analysis in personality disorders research: Modern issues and illustrations of practical recommendations. *Personality Disorders, 14*(1), 105–117. https://doi.org/10.1037/per0000581

Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2015). Methodology, ontology, and interventionism. *Synthese, 192*(11), 3577–3599. https://doi.org/10.1007/s11229-014-0479-1

Woodward, J. (2016). The problem of variable choice. *Synthese, 193*(4), 1047–1072. https://doi.org/10.1007/s11229-015-0810-5

Woodward, J. (2021). Downward causation defended. In J. Voosholz, & M. Gabriel (Eds.), *Top-Down causation and emergence* (pp. 217–251). Springer International Publishing.

Yeung, E. K. L., Apperly, I. A., & Devine, R. T. (2024). Measures of individual differences in adult theory of mind: A systematic review. *Neuroscience & Biobehavioral Reviews, 157*, Article 105481. https://doi.org/10.1016/j.neubiorev.2023.105481