Truth and Evidence in Validity Theory

Denny Borsboom
University of Amsterdam

Keith A. Markus
John Jay College of Criminal Justice of The City University of New York

According to Kane (this issue), “the validity of a proposed interpretation or use depends on how well the evidence supports” the claims being made. Because truth and evidence are distinct, this means that the validity of a test score interpretation could be high even though the interpretation is false. As an illustration, we discuss the case of phlogiston measurement as it existed in the 18th century. At face value, Kane’s theory would seem to imply that interpretations of phlogiston measurement were valid in the 18th century (because the evidence for them was strong), even though amounts of phlogiston do not exist and hence cannot be measured. We suggest that this neglects an important aspect of validity and suggest various ways in which Kane’s theory could meet this challenge.

We welcome Michael Kane’s article updating, extending, and further developing his approach to test validity. We choose to focus our comment on a single issue that we consider fundamental to test validity and on which Kane’s account remains ambiguous: How does the truth of the conclusion of the validity argument fit into validity in Kane’s approach?

Modern test validity theory contains a dialectic between two metaphors. The first metaphor is the mechanical metaphor in which a test is a machine that measures and is valid if it measures what it is intended to measure. For example, one can think of an alarm clock as a test that is reliable if it sounds consistently at some specific time and valid if it sounds consistently at the intended time. The second is the argument metaphor in which an interpretation is valid if it follows from test scores by way of a valid argument. Thus the conclusion that the time has come to get up follows appropriately from the sounding of the alarm if the inference is valid. The mechanical metaphor emphasizes the link between test validity to the truth of conclusions drawn from test scores (Borsboom & Mellenbergh, 2007; Borsboom, Mellenbergh, & Van Heerden, 2004). The argument metaphor emphasizes evidence to support the inference.

One way to frame this issue is in terms of the traditional analysis of knowledge as justified true belief (Shope, 1983). The mechanical metaphor emphasizes true belief, whereas the argument metaphor emphasizes justified belief. The argument metaphor begins with the prototype of a deductive argument in which the conclusion must hold true if the premises hold true. In an ideal Cartesian tree of knowledge, only certainly true beliefs would be admitted as premises, and thus conclusions would follow with certainty. Thus truth and justification hold together. However, even formally valid deductive arguments allow for some slippage: A valid argument can lead to a false
conclusion if it begins with false premises. Thus what is required is not just validity but soundness of the argument (validity plus true premises).

As Kane emphasizes, however, science does not typically proceed by deductive argument. Science proceeds by ampliative arguments that draw conclusions that go beyond what is made certain by their premises. Extending the metaphor, such inferences can be deemed valid if they tend toward true conclusions without guaranteeing them. A standard example is inductive inference based on a bad sample, in which a sample of all red spheres drawn from an urn containing 50% red spheres leads by justifiable inference to a false conclusion about the proportion of red spheres in the urn. As such, truth and justification can come apart, at least in the short run. Messick (1989) gave the example of test behaviors that may be adaptive in some contexts but not others being wrongly interpreted as universally adaptive (e.g., consistency) or nonadaptive (e.g., rigidity). An example like this holds even in the long run, because one can collect endless empirical evidence that the scores align to the construct without detecting the misinterpretation applied to both the construct and the scores when interpreting the research evidence. The informal arguments emphasized by Kane are clearly ampliative rather than demonstrative, widening the potential gap between justified belief and true belief.

It is not our intent to advocate for one metaphor over the other, as we think that both involve important questions in test validity theory (Markus & Borsboom, in press). Instead, we wish to focus on a highly specific question: How, if at all, is the truth of the conclusions drawn from test scores incorporated into Kane’s approach to validity? To put a finer point on it, does Kane’s approach provide a framework that makes it possible to represent a situation in which the best available evidence leads to a false conclusion? If not, is this a deficiency of Kane’s theoretical framework or a deliberate assumption that justification and truth cannot come out of alignment?

This is not merely academic curiosity, in our view, because it has practical implications for test validation. If test validity theory emphasizes justified belief to the exclusion of true belief, validation may become an end in itself rather than a means to an end. In such a case, one constructs an argument to support a test score interpretation simply because one wants to support that interpretation. In this case, validity arguments risk becoming akin to arguments for conspiracy theories: they never fail to support these theories simply because they are explicitly designed to do so. In our view, however, one constructs and evaluates a validity argument as a means to an end—namely, because one wants to arrive at a better understanding of how well the test is functioning. This, however, requires an account of validity that incorporates both justified belief and true belief as distinct elements. The question before us is whether Kane’s approach has sufficient conceptual resources to achieve this.

Phlogiston: A Clear Test Case

Cronbach (1971) emphasized explanation, which assumes truth, and Messick (1989) presented an elaborate theory of fallibilism. However, in Kane’s presentation of the argument-based approach, the emphasis is entirely on justification to the exclusion of truth. As Kane puts it (this issue, p. 1), “the validity of a proposed
interpretation or use depends on how well the evidence supports the claims being made.” Rather than link the argument to justification and the concluding interpretation to truth, Kane links the interpretation itself to justification. Thus, taking Kane’s statements at face value suggests that the validity of test-score interpretations (not just the arguments supporting them) is essentially independent of their truth. This makes validity entirely a time-dependent concept that is relative to scientists’ evidence and theories. Without the qualifications of earlier theorists, Kane asserts “Validity . . . may change over time, as the interpretations/uses develop, and as new evidence accumulates” (p. 3). The time-independent element of validity involving truth seems to get lost.

To make the problem clear, it is useful to briefly illustrate it with a case where evidence for a long time supported a wholly incorrect interpretation of measurement outcomes: the measurement of phlogiston (see Borsboom, Cramer, Kievit, Zand Scholten, & Franic, 2009, for a detailed discussion). The theory of phlogiston (“fire-stuff”) posited the existence of a substance contained by flammable materials that was emitted in the form of fire when these materials were heated. In the 18th century, scholars measured the amount of phlogiston that a piece of material contained by subtracting the weight of the material after burning from the original weight of the material: the difference was thought to equal the relevant amount of phlogiston. Call this test-score interpretation Interpretation P: “the weight of a substance before burning minus the weight of the same substance after burning equals the amount of phlogiston the material contained.”

Support for Interpretation P existed in the form of a quite impressive theory on the nature of burning as phlogiston emission. For instance, the theory of phlogiston could explain why some materials burned while others did not (they did not contain phlogiston), why materials that do not burn (e.g., iron) do not lose weight when heated (they do not emit phlogiston), why a burning candle dies out if there is no fresh air supply (the air becomes saturated with phlogiston), and so forth. Thus, until the end of the 18th century when Lavoisier refuted the theory and showed that burning is a chemical reaction, measurement interpretations in terms of phlogiston enjoyed considerable support.

The example provides an interesting test case for validity theories because it clearly shows how truth and evidence can come apart in test-score interpretations: Interpretation P was never true, but it was supported by significant amounts of evidence and strong arguments. Its negation, not-P, was always true but was not supported by evidence before Lavoisier entered the scene. If we take Kane literally, then the phrase “the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made” would seem to imply that Interpretation P had high validity in, say, 1730. Thus, if it had been applied by the proponents of phlogiston theory in the early 18th century, Kane’s theory of validity would have led to the acceptance of phlogiston measurement as valid.

The question we would like to pose to Kane is whether this is indeed a correct reading of his theory. If so, how would he evaluate the phlogiston measurement example? As we see it, the inference may have been valid, but the interpretation of measurement outcomes was not. In fact, it seems to us that phlogiston measurement is a textbook example of invalidity in test-score interpretation and that a good theory
of test validity should somehow accommodate this. Regardless of how subtly one constructs the relation between test score interpretations and reality, it would seem that any theory of validity should deem the interpretation invalid.

Possible Strategies for Dealing with the Phlogiston Case

We see several possible responses that Kane could provide in response to the phlogiston case. First, he could bite the bullet. That is, Kane could accept that his theory would have supported Interpretation P in the 18th century (not just the inference). This would imply that Kane’s theory emphasizes justified belief to the exclusion of true belief. In our view, this would be a heavy price to pay, but Kane could choose to do so. Also, if this should be the preferred route, then it would seem that Kane’s theory is incomplete and requires a supplement outside of what it labels validity to deal with the relation between measurement interpretations and the world (i.e., true belief about test scores). Even a thoroughgoing pragmatist such as Rorty (2000) recognizes the need for a minimal notion of truth to support fallibilism about belief.

Second, Kane could attempt to show that the argument does not actually stick. This would require that Kane set up an argument to the effect that his theory would not actually have granted validity to Interpretation P in 1730. For instance, Kane could attempt to furnish his approach with the room for fallibilism that all theories of truth require. This could for instance be done by including something like an “ultimate argument” into his argument-based approach; for instance, by making validity relative to the argument that rational scientists would arrive at, should they continue their investigations for an infinity of time. That way, Interpretation P could be valid to the observers in the 18th century but invalid with respect to the ultimate state of scientific knowledge. Such approaches to truth face many challenges, but the approach would allow Kane to incorporate both justified belief and true belief into his approach while at the same time keeping the emphasis on justified belief.

Third, Kane could attempt to show that even though his theory is unable to rule correctly about Interpretation P, other theories cannot do so either. This strategy would come down to showing that, even though it seems that a theory of validity like that of Borsboom et al. (2004) correctly rules Interpretation P to be invalid (because phlogiston does not exist and thus cannot produce measurement outcomes, so that a sufficiently strong causal reading cannot be given; Markus, 2004, 2008), this is not actually so. Alternatively, Kane might argue that alternative approaches shortchange justified belief the same way that his approach shortchanges true belief and that shortchanging true belief is the lesser of two evils. The challenge would be to show why it is necessary on such a view to accept shortchanging one or the other.

However Kane chooses to treat the issue, we think that any validity theory should deal with the relation between evidence and truth, however difficult that may be in psychological and educational testing (Markus & Borsboom, in press). In our view, a theory of validity that exclusively deals with how to organize evidence and justify decisions misses an essential psychometric aspect of validity and is unnecessarily impoverished.
References


Authors

DENNY BORSBOOM is Professor of Psychological Methods at the Department of Psychology of the University of Amsterdam, Weesperplein 4, 1018 XA Amsterdam, The Netherlands; dennyborsboom@gmail.com. His primary interests include psychometrics, philosophy of science, and network modeling.

KEITH A. MARKUS is Professor of Psychology at John Jay College of Criminal Justice of The City University of New York, Psychology Department, 524 West 59th Street, New York, NY 10019; kmarkus@aol.com. His primary interests include test validity, causal explanation and inference, and program evaluation.