


ARTICLE

# Measuring the Nonexistent: Validity before Measurement

Kino Zhao 

Simon Fraser University, British Columbia, Canada  
Email: [kino\\_zhao@sfu.ca](mailto:kino_zhao@sfu.ca)

(Received 21 February 2021; revised 27 December 2022; accepted 06 January 2023; first published online 17 February 2023)

## Abstract

This article examines the role existence plays in measurement validity. I argue that existing popular theories of measurement and of validity follow a correspondence framework, which starts by assuming that an entity exists in the real world with certain properties that allow it to be measurable. Drawing on literature from the sociology of measurement, I show that the correspondence framework faces several theoretical and practical challenges. I suggest the validity-first framework of measurement, which starts with a practice-based validation process as the basis for a measurement theory and only posits objective existence when it is scientifically useful to do so.

In his article boldly titled “General Intelligence, Objectively Determined and Measured,” Spearman (1904) presents a series of experimental data on children’s judgments of pitch, brightness, and weight and argues that because children who judge accurately in one area tend to judge accurately in others, this is conclusive evidence that intelligence is an innate characteristic of people that psychologists can and should study. The statistical method he developed for this project, factor analysis, was later applied in another area of psychology to make a similar point. Citing various studies on the cross-time stability and predictive validity of the five-factor personality model, McCrae and John (1992) declare the model to be “a basic discovery of personality psychology—core knowledge upon which other findings can be built” (see also Costa and McCrae 1992).

Both of these cases employ an inference to the best explanation (IBE) style of reasoning to argue for the realism of some psychological entity—because the reality of the proposed construct is the best explanation of its predictive capabilities, accurate prediction is evidence for the existence of the construct. Citing its numerous predictive successes, McCrae and John (1992) write that without positing the reality

of the Big Five theory of personality, “it is difficult to understand how cognitive fictions can explain real-life outcomes” (193).

In other words, the debate about the reality of the construct turned on claims about the validity (in this case, predictive validity) of the tests. Realists like Spearman and McCrae argue that the validity (predictive, criterion, and construct) of concepts like intelligence or the Big Five demands explanation and that the best explanation on offer is that these concepts have some sort of psychological, if not biological, reality. Critics have also often accepted this argumentative strategy, focusing instead on challenging the particular validity claims made about these concepts or proposing equally good explanations that, although challenging a specific theory of reality (e.g., a five-factor instead of a two-factor model; see Eysenck 1991; Boyle 2008), remain committed to the general picture that validity ought to have some indicative power to realism.

The connection between measurement validity and scientific realism has been a recurrent theme in validity theory. Historically, some have opted for a “thin” conception of validity, where “a test is valid for anything with which it correlates” (Guilford 1946), and “[t]o claim that a test measures anything over and above its criterion is pure speculation” (Anastasi 1950), whereas others insist on having robust commitments to realism, taking validity to track “the degree to which [the test] measures some trait which really exists in some sense” (Loevinger 1957). More recently, (Borsboom et al. 2004) have argued that “[t]he attribute to which the psychologist refers must exist in reality; otherwise, the test cannot possibly be valid for measuring that attribute,” whereas (Kane 2013b) protests that “[i]n many testing situations (including most high-stakes contexts), talk of Truth seems hollow.”

Whether measurement validity entails scientific realism has important philosophical as well as scientific consequences. This is especially true in the social sciences,<sup>1</sup> where people are generally hesitant regarding realist commitments about constructs. Even if we are not reductionist physicalists, calling a construct “real” has scientific salience. A real construct can be studied across fields and contexts and can play certain causal roles. If general intelligence is a real psychological concept, then it makes sense to ask which brain structure is associated with high intelligence or what the genetic makeup of intelligence is. If it is not, then we know that any explanatory story citing intelligence as a cause must be a work in progress because “intelligence” would merely be a placeholder for something (possibly a number of different things) more fundamental.

The present article challenges the belief that because validity claims presuppose the existence of the construct under measure, we can infer realism about the construct from the validity of the tests through IBE. In particular, I argue that validity claims need to fulfill important practical roles, which makes them unsuitable to also carry the kind of epistemic burden an IBE for realism demands. Instead, we should develop measurement theories that do not take realism to be a precondition for successful measurement.

---

<sup>1</sup> Generally speaking, I do not subscribe to a well-defined “social” versus “natural” distinction about the sciences. However, many, although not all, theories and examples I draw upon in this article are from fields typically classified as social sciences (notably, psychology, education, anthropology, and sociology). I therefore use the term *social sciences* to refer to these fields, but I do not think that any observations I make are unique to them.

This article is organized as follows. Section 1 argues that currently popular theories of measurement—the representational theory of measurement (RTM), classical test theory (CTT), and item response theory (IRT)—and the validity theories that go along with them all follow a *correspondence framework of measurement*, according to which to measure is to accurately capture the properties of something that objectively exists in the world. According to this framework, because existence is the precondition for measurement, which is the precondition for validity, validity claims need to be explained by existence claims, allowing for IBE reasoning of the form mentioned earlier.

Section 2 challenges the adequacy of this framework in its description of scientific practice. Not only do scientists measure constructs whose reality they question, but they also assess the validity of measuring instruments without referencing the underlying reality of the constructs. Because the correspondence framework cannot account for these measurement practices, it cannot provide theoretical guidance for them. I argue that this is because the correspondence framework takes measurement to be a passive, descriptive project and ignores its creative potential.

Section 3 sketches an alternative framework for understanding measurement and validity. Drawing on recent developments in the “argument-based approach to validity” (Kane 2013a), I develop a novel framework for understanding measurement that puts validity judgment at the foundation for measurement theories and existence claims. I call this the *validity-first approach*. Section 4 concludes.

## 1. The correspondence framework

The most popular measurement theory in philosophy is the representational theory of measurement (RTM). According to this view, a quantity is measurable just in case there exists an isomorphism between a construct and the number system. In what follows, I provide a brief history of the development of the RTM and discuss the kind of realism I take this theory to be committed to. I then turn to measurement and validity theories in psychology and consider whether operationalism prevents this kind of realism.

The initial motivation for what later became the RTM was not to use numbers to represent properties but to use properties to define numbers. When von Helmholtz wrote *Zählen und Messen* ([1887] 1930), his aim was to found arithmetic by axiomatizing counting. To him, therefore, the measurement target (discrete objects) is more fundamental than the numbers we use to represent it because, coming from a Kantian perspective, the measurement target is more empirically accessible than numbers.

Not all empirically accessible properties are representable by numbers, however. Helmholtz lists the usual suspects—length and weight—as examples of attributes that share enough structural similarities with numbers such that they can be “counted” (measured) in the same way as the quantity of discrete objects. Other attributes, such as pain and pleasure, cannot be represented in this way. Whether an attribute has enough structural features to be measurable is something we discover in the world. The fact that length and weight are representable by numbers by way of an isomorphism is true regardless of whether we have ever tried to measure them. The fact that pleasure and pain lack key structural features, which prevents them from being (fully) representable by numbers, is true even if the psychologists do not like it.

The subsequent debate between the physicists and the psychologists<sup>2</sup> on the measurability of sensations likewise did not turn on the properties of sensations, which were taken as given in the world, but on the definition of measurability. Stevens's (1946) insight is that numbers can carry partial information about an attribute without full-blown isomorphism. His theory of scales differentiates between the kinds of structural information that numbers can carry and that warrant some but not other inferences. It is true that the sensation of brightness does not admit a concatenation procedure and therefore does not obey the additivity axiom of numbers, but we can still measure it on an ordinal scale; we just have to be careful not to use addition in our subsequent statistical analysis.

Although Stevens (1946) makes assertions like “measurement . . . is defined as the assignment of numerals to objects or events according to rules,” his operationalism is not as antirealist as that of others in his circle, such as Boring (1923) when he claims that “intelligence is simply what the tests of intelligence test” (Hardcastle 1995). Stevens, as well as his successors, such as Suppes and Zinnes (1963), still operate on the assumption that the measurement target is given in the world, and the job of the measurement theorist is to discover its structural properties and represent them with accuracy.

Talking about success in measurement in terms of representational accuracy immediately brings up two philosophical issues. The first is metaphysical: What is the thing that is accurately (or inaccurately) represented? As explained previously, the RTM operates on the assumption that there is a well-defined, objectively existing attribute that we are trying to represent through measurement. In their RTM-inspired theory of measurement, Bradburn et al. (2017) argue that measurement can be seen as a three-step process. First, we define the target concept (characterization); next, we define a metric system that represents it (representation); finally, we formulate rules for applying the metric system (procedures). According to this view, concepts that cannot be characterized with enough clarity cannot be candidates for measurement.

The second issue with evaluating measurement in terms of accuracy is epistemological: How do we know if a representation is accurate? In the traditional RTM, representation is accomplished through isomorphism, which is proven between axiomatizations of the target concept and the metric system. Whether the representation is accurate, therefore, depends on whether the axiomatization is faithful, which in turn depends on how much we know about the behavior of the target concept. This is easy to determine in the case of mesoscale, observable attributes that can be easily manipulated, such as the lengths of rods. It is much less feasible when the target concept cannot be accessed through measurement-independent ways.

The difficulty of determining whether our metric system accurately corresponds with an unobservable target has been termed the *problem of nomic measurement* by Chang (2004). The worry is that if the only way to access a construct is through measurement, we have no epistemic foundation on which to calibrate that measurement. One common coping strategy is to invoke a kind of robustness reasoning across multiple forms of measurement—even if none of them is

---

<sup>2</sup> See Campbell and Jeffreys (1938) for arguments against and Stevens (1946, 1968) for arguments in favor of the measurability of sensations. See Michell (1999) for a historical overview.

independently calibratable, if they all give the same output, perhaps that is evidence that they are all accurate. But this strategy runs into another problem, one of quantity individuation identified by Tal (2019). Tal argues that when multiple measurement procedures produce different outcomes, it is often underdetermined whether this is evidence for the inaccuracies of the procedures or that they actually measure different constructs.

Both Chang and Tal reject a kind of measurement foundationalism whereby the (objectively existing, stable) features of the measurement target serve as the foundation that informs the construction of the measuring instrument, which in turn dictates the interpretation of its results. The RTM subscribes to this kind of foundationalism. So, as I shall argue later on, do other existing measurement and validity theories. However, although I will end up rejecting this kind of foundationalism like Chang and Tal, I will not endorse coherentism like they do. Instead, I will propose a different kind of foundationalism that makes validity theory the foundation for measurement theory and avoids the aforementioned problems.

It is worth noting<sup>3</sup> that some (Bacelli 2020; Vessonen 2021) have argued that the RTM should not be understood as a theory of measurement but instead as a theory of measurability. It is, therefore, not a fair criticism to point out that the RTM fails to describe how measurement actually occurs in science. Adopting this (in my view, convincing) perspective on the RTM, the present article can be seen as challenging the RTM definition of measurability.

Two other popular measurement theories exist in addition to the RTM: classical test theory (CTT) and item response theory (IRT; also called *latent variable theory*). Traditionally, CTT has been associated with operationalism, which is relatively weak in its realist commitments when compared with the RTM and IRT. Nevertheless, the core idea of CTT is that a measurement result (“observed score”) is composed of two parts, the true score and a random error, where the true score is the measurement-independent reality we are trying to study. CTT has been criticized along lines similar to Tal’s, where critics complain that there is no interpretation of the true score that is both epistemically justified and scientifically useful.<sup>4</sup>

In a sense, IRT avoids the problem of quantity individuation by starting with the assumption that a collection of instruments measures the same small number of latent variables, which are causally responsible for the observed measurement results. Once this assumption is in place, IRT helps us determine the degree to which each test item measures each latent variable. It is the causal relationship between the latent variable and the test item that grounds measurement, and it is a fact in the world whether this relationship exists.

Insofar as measurement is seen as a descriptive project, this correspondence picture seems inevitable.<sup>5</sup> Indeed, all three theories of measurement just reviewed rely on the assumption that the measurement target has a measurement-independent existence, and to measure is to capture its properties through the establishment of a

<sup>3</sup> I thank a reviewer for both the point and the references.

<sup>4</sup> For a review of this debate, see Borsboom (2005).

<sup>5</sup> A reviewer expressed the sentiment, which is perhaps widely held, that the correspondence picture is definitional of measurement. It is precisely this sentiment that I seek to challenge in the present article.

relationship, representational in the case of the RTM and causal in the case of IRT and CTT, between a target and the results. This naturally entails the concept of validity as the evidenced success of such an establishment.

Validity theory, as it currently stands, is composed of three aspects:<sup>6</sup> content-oriented validity, criterion-oriented validity, and construct validity. Very briefly, content-oriented validity consists of seeing if the wording of test questions sounds like reasonable descriptions of the desired attitudes. Criterion-oriented validity involves correlating results from the present test with other observables expected to correlate with the construct under measure, such as predicted behaviors. Finally, construct validity involves building an elaborate theory of the behavioral implications of a construct and testing it through extensive research.

The extent to which these validity theories rely on the realist assumption is not always clear. Stemming from education research, there is a significant practical aspect to the problems faced by validity theorists. Many are tempted to adopt a “thin” notion of validity, where a test is called valid if it correlates relatively strongly with some plausible criteria. This makes validity claims easier to come by, but test results become less scientifically useful. On the other hand, insisting on a “thick” notion of validity, where a test is only valid if it is capable of truthfully describing the external world, leads to an inconveniently few tests we want to use qualifying as practically validatable.

In fact, the tension between theoretical strengths and practical limitations has been a recurrent struggle within validity theory. When Cronbach and Meehl (1955) proposed their influential theory of construct validation, many testers claimed to have been convinced—they understood that test validation was a process by which researchers gathered behavioral data in order to empirically confirm a “strong theory” about the objectively existing construct, which was the target of measurement. Heavily influenced by logical positivism, the idea was that a construct is a theoretical entity that connects to the world through a “nomological network,” which ultimately exhausts into a set of observational sentences. To validate a construct is to confirm (or refute) this strong theory by providing observational evidence for or against it. Because theory confirmation is a difficult and possibly never-ending process, so is validation.

Although the construct-validity program gave validation a long-sought-after theoretical framework that is philosophically well grounded and scientifically intuitive, it also highlighted the core conflict between the theoretical demands and practical challenges of validation in full and unavoidable terms. Test developers realized that very few constructs had the kind of strong theories Cronbach and Meehl envisioned, and they often needed to make test-use decisions when the evidence was nowhere near theory-confirmation level. Only a couple of decades later, Cronbach (1980) complained that a typical validation report in the literature had become “an unordered array of correlations with miscellaneous other tests and

---

<sup>6</sup> The division can be traced back to as early as when construct validity was first proposed in “Technical Recommendations for Psychological Tests and Diagnostic Techniques” (American Psychological Association 1954), but the relationship among the divisions is not always clear. For example, Cronbach and Meehl (1955) believe that construct validity should replace the other two, whereas Messick (1989) argues that different circumstances call for different types of validation. The most recent *Standards for Educational and Psychological Testing* publication (American Educational Research Association et al. 2014) considers them as complementary sources of evidence to be used in the argument-based approach to validation.

demographic variables. Some of these facts bear on construct validity, but a coordinated argument is missing.”

We can see another example of the tension between realism-assuming validity theories (notably, the construct-validity theory) and practical demands of validation in the debate around whether the context of test use should affect the assessment of validity. Shepard (1997) considers a case where pre-med students prioritize science classes over humanities as a way to increase Medical College Admission Test (MCAT) scores, thus making the MCAT no longer an adequate test for identifying students who are more likely to succeed in medical school. The act of using the test has changed the usefulness of its results. As Shepard points out, if we take the usefulness of a test's results to be indicative of, but not identical to, its validity, then we must conclude that either (1) the validity of the MCAT changed in response to students' choice of classes, or (2) our earlier assessment of the validity of the MCAT was mistaken. The problem with (1) is that according to construct validity and the general correspondence framework of measurement I have been sketching, the validity of a measurement consists of the quality of the relationship between the measurement result and the objectively existing construct, and it is unclear how this relationship should have changed in the MCAT example. Option (2) is even less desirable because almost all (external) validation procedures come down to claims of usefulness. If the MCAT was useful before students' change of behavior, then there is little reason to retroactively deny the earlier claim of validity. Shepard uses this example to motivate a thin view of validity, where validity claims do not rely on the successful discovery of a specific way that the world is and thus can change as contexts change.

Ultimately, the choice between thick and thin views of validity is a matter of preference: Would we rather hand out lots of validity claims that don't mean very much, or would we rather hold validity as the ultimate stamp of approval even if it means that large parts of science need to make do with unvalidated measurements? Historically, measurement theorists have mostly opted for thick views of validity that are difficult to obtain but carry substantive theoretical weight once established. From the perspective of the correspondence framework of measurement, a measurement is valid if the measurement results have accurately captured the objective properties of a construct that really exists in the world. A validity claim about a measurement procedure, therefore, warrants a corresponding realism claim about the target construct.

The present article can be seen as advocating for a serious entertainment of the thin option. The theoretical strength and practical limitations of the thick-construct-validity project have been extensively studied by the measurement community, but theorists are, understandably, resistant to the suggestion that we should weaken the epistemic power of our science to fit practical constraints. It is part of my goal in this article to show that taking the thin conception of validity does not need to be a theoretical concession; it may be philosophically and scientifically fruitful enough to be a genuine competitor to the standard conception. In the next section, I review some reasons for adopting the thin view of validity that go beyond its practical convenience.

## 2. Measuring the nonexistent

The kind of measurement cases I will focus on are ones that start with a relatively clearly defined practical context of a test without any accompanying substantive



theory of a construct. For example, a company might use a simple customer satisfaction survey to determine whether changing its service in certain ways will lead to an increase in satisfaction, consumer loyalty, and ultimately, profit. In cases like this,<sup>7</sup> it is usually fairly straightforward to determine whether a test has succeeded in being useful—it is successful if the increase in profit coheres with what the test developers promise. It is much trickier to claim that this success is the result of accurately capturing the objectively existing degree of satisfaction each customer feels about a product. More importantly, it is unclear why the lack of such a theory of the construct of customer satisfaction should be an obstacle to calling some of these surveys good or “valid.” In other words, the reality of the construct is an unnecessary intermediary between the test context and claims of its validity.

The correspondence framework relies on this intermediary. Because the correspondence framework takes measurement to be a *descriptive* project, the success of measurement (as shown through validation) naturally implies the success of the description, which in turn implies the existence of the thing being described. Validity needs explanation. In a descriptive project, the best explanation for validity is the reality of the thing being described.

In what follows, I challenge the view of measurement as a merely descriptive project. Drawing on historical and anthropological studies of measurement, I show how measurement often changes our conceptualization of the world—and consequently, the world itself—in profound ways. I will then argue how this new view of measurement does not see validity as something needing to be explained by realism.

There are, roughly speaking, three kinds<sup>8</sup> of measurement–world interactions I will highlight. The first is when “merely” arbitrary choices about measurement procedures change which parts of the world are open to scientific study and in what ways. The second is when choices made during measurement build into the foundations of our theoretical understanding of the phenomena. The third is when the act of measurement literally causes the world to change in response.

Despite the widespread acknowledgment of the inevitability of arbitrary procedural choices during measurement, it is difficult to appreciate the extent of their influence on measurement results. For example, as Porter (1996) observes,

In principle, the population of a country is a relatively unproblematical number. But it is not fully determined by the distribution of bodies over a landscape. First a decision must be reached about how to count tourists, legal and illegal aliens, military personnel, and persons with more than one residence or multiple citizenship.

There is a sense in which it doesn’t matter which way we choose, as long as we take care to be consistent across time. But consistency assumes a certain level of

---

<sup>7</sup> A reviewer has expressed the sentiment that cases like this are examples of misusing the concept of “measurement” and are not worth being taken seriously. My view is that these application contexts are sufficiently ubiquitous to at least warrant serious examination. As I seek to redefine “measurement” as a scientific concept, I will proceed with a theoretical agnosticism about the actual merits of these apparent claims of measurement.

<sup>8</sup> I do not see them as differing in kind but rather as differing in degree. Nothing I say will hinge on the nature of their difference.



retainment of auxiliary information that doesn't always occur. If I'm trying to measure population growth, then including legal aliens but not military personnel seems harmless as long as this is done consistently across time and there isn't a sudden surge in enlistment. But the judgment that this is done consistently can only be formed if there is some memory of how it was done in the past. Because these arbitrary choices are often dismissed as theoretically uninteresting, they are seldomly recorded. For example, the National Comorbidity Survey (NCS) "did not include supplemental samples of other institutional populations (e.g., prisons, hospitals, nursing homes) or of the homeless population" (Mickelson et al. 1997) for cost reasons, but it also did not include a sketch of the kind of institution that would fall under this category.

The decision is arbitrary in the sense that choosing one way or another (often) does not have (immediate) consequences for the data-based theorizing at hand, but the innocence of these decisions also shields them from critical scrutiny. Although one could, in principle, contact the NCS surveyors to get a more detailed picture of all the judgment calls they've made in the survey, in practice, there is rarely an incentive for that. As data sets age, they become more entrenched and less challenged, which, paradoxically, makes it difficult to assess the true extent of their innocence.

Merry (2016) calls this phenomenon *data inertia*. Because gathering data is expensive, organizations prefer to either repurpose old data or, when they must generate new data, minimally adapt entrenched measurement procedures. Even when there is a genuine effort to develop new measurements, old instruments and data are often taken to be starting points at first and validation anchors afterward. In other words, entrenched ways of measurement, however arbitrary they may have been at their creation, often end up exerting disproportional influence over subsequent measurement efforts.

Merry argues that phenomena like data inertia and expertise inertia, where experts who were there at the beginning of the project exert disproportional influence over later development, make international collaborations on measurement less democratic than they advertise. Once the initial attempt is made, it defines the key parameters of subsequent development. Deviations need to be justified, whereas conformities do not. Challenges are expected to be posed with existing terms and concepts before they are taken seriously. In Merry's words (2016):

The expertise of these actors and the availability of data shape the way they categorize and analyze information to develop an indicator. The politics of indicators are visible in the way categories are constructed, decisions are made about what to count, and concepts are defined as measurable. The knowledge they provide is inevitably interpreted through their expertise and experience.

To be clear, the worry is not that we have reasons to believe that some entrenched framework of measuring is flawed. The worry is that the cause for its entrenchment is not truth tracking and that once a framework is entrenched, it is difficult to assess its real merit. In other words, if some frameworks are in fact flawed, we would never know.

We may call data inertia an example of an epistemic consequence of practical constraints. Theoretically speaking, nothing prevents us from constructing completely new tests of the same phenomenon each time we need to measure it. Practically, however, doing so would be deemed a waste of time and resources. Moreover, because measurement results are usually only interpretable in reference to the measuring instrument, having multiple radically different instruments also harms the usability of results. All of these practical reasons discourage experimentation.

In fact, Porter (1996) argues that the fact that we tend to overestimate the strength of the measurement–nature correspondence necessary for measurement success is exactly why numerical measurement is so ubiquitously adopted in social settings. Quantification, argues Porter, is often valued for what it has to leave out as much as for what it is capable of capturing. The process of taking a diverse set of phenomena, imposing a kind of quantitative uniformity onto them, and making it look like the decision is objective and therefore fair is an act of political control that is very often consciously done by measurement agencies. More recently, John (2022) has similarly observed the use of spuriously precise numbers as a mechanism for behavioral manipulation.<sup>9</sup>

Not only is the manipulative use of numbers often successful, but measurement can also change how we relate to the world in ways that hide the fact that their descriptive success was not caused by their correspondence with nature. One way that this can happen is when the adoption of a measurement framework changes how we understand the world.

For example, Siegel (1994) has documented the conceptual change during the late 19th century in understanding women’s household labor as a kind of work. The initial movement was motivated by a legal demand that wives should have a share of the domestic property. Because property rights were tied to labor contribution, the issue naturally fell on the question of whether a wife made a labor contribution to her family. That is, the question was essentially about how we should measure labor contributions: Should we count household chores or not?

It is not the kind of question that could be answered rightly or wrongly in the same way that a question about which rod is longer could be. But it is also not the kind of question that is completely arbitrary. Indeed, through years of difficult feminist work, the question has been given an answer that we now commonly think of as correct: full-time housewives do make labor contributions to the household. Providing this answer not only resolves the original measurement question, however; it also shapes our understanding of what labor is and what making a contribution to the household can look like.

However, not recognizing housewives’ labor was not a mistaken assessment of reality, like failing to properly count the number of people in a room is. Switching from the old way of measuring labor to the new way is not the same as replacing the inaccuracies of an old understanding with accurate details. To say that the present way of measuring labor is the right way because it corresponds with facts in the world is to overlook the conceptual revolution that was necessary to get us here. In Merry’s

---

<sup>9</sup> I thank a reviewer for pointing me to this reference.

words, “those who create indicators aspire to measure the world but, in practice, create the world they are measuring” (Merry 2016).

Finally, how we choose to measure the world can not only change how we understand the world; it can also cause the world to literally change. To give an example that is outside the stereotypical social sciences, Scott (1996) has observed how standardized measurement practices in forestry change people’s relationship with forests:

The achievement of German forestry science in standardizing techniques to calculate the sustainable yield of commercial timber and, hence, revenue, was impressive enough. What is decisive for our purposes, however, is the next logical step in forest management. That step was to attempt to create, through careful seeding, planting, and cutting, a forest that was easier for state foresters to count, manipulate, measure, and assess. The fact is that forest science and geometry, backed by state power, had the capacity to transform the real, disorderly, chaotic forest so that it more closely resembled the administrative grid of its techniques.

That is, the fruitfulness of the imposed measurement system has caused a change, whereby nature is intentionally and explicitly manipulated to conform to the measurement system. Although it is true that a valid measurement result provides an accurate (in the correspondence sense) representation of reality, this is not because the measuring instrument has succeeded in its descriptive goal. It is the exact opposite: the world has been bent to the prescriptive power of the measuring instrument. The correspondence framework, which treats valid measurement as an accurate description of the measurement-independent reality, obscures the creative dynamics often present between the world and our attempts at making sense of it.

To be clear, neither I nor these authors are suggesting that measurement creates concepts out of thin air and imposes them onto the world against its will. What I am arguing is that the structural features of successful measurement are often not chosen to best reflect nature but are selected for a variety of practical and political reasons. Once selected, it is difficult not to see the world through the carefully crafted lens that is the measurement system. As theorists, we should therefore be especially careful when making inferences about the structure of the world on the grounds of measurement results alone. In the next section, I provide a sketch of how I believe we should approach measurement theory instead.

### 3. A validity-first framework of measurement

If measurement procedures are not developed for their ability to accurately capture objective features in the world, then how can we judge measurement quality? In fact, once we abandon the correspondence framework, judging the quality of a measure is often easier than theorizing about why a measure achieves this quality.

As already mentioned in section 1, the construct-validity program of Cronbach and Meehl (1955) enjoyed widespread celebration for its theoretical virtue. It is still the most acknowledged validity theory today, even by philosophers who have pointed to its deficiencies (e.g., Alexandrova and Haybron 2016; Stone 2019). Nevertheless,

by 1980, Cronbach had already conceded to testers who were prevented from following the spirit of the program by practical difficulties.

In the fourth edition of *Educational Measurement* (Brennan 2006), a self-described “bible in its field” and recurrent publication sponsored by the American Council on Education and the National Council on Measurement in Education, Kane advocates understanding validity as a relationship between an interpretation of the test scores and a specific use context. In Kane’s words, “to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use. The evidence needed for validation necessarily depends on the claims being made” (Brennan 2006).

This view has been called the argument-based conception of validity. It was adopted by the 2014 edition of the *Standards for Educational and Psychological Testing*, a joint publication by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, according to which:

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. . . . Statements about validity should refer to particular interpretations for specified uses. It is incorrect to use the unqualified phrase “the validity of the test.”

In other words, the *Standards* advocates viewing validity not as a property of the test—that it accurately describes the world—but as the success of test results when applied to a specific context. Because tests are almost always developed for a specific practical goal, there usually exist some concrete criteria by which funding agencies can judge if their goal has been achieved. For example, an academic entrance exam serves its purpose if students who do well on this exam tend to do well academically after admission.

I have been calling this the “thin” conception of validity. Unlike the thick-construct-validity program, the thin conception makes the validity label easy to apply while also taking away its theoretical substance. We are no longer justified in inferring anything straightforward about the world or about the test from a claim of validity alone because claims of validity are always relativized to specific interpretations and use contexts.

For those who see measurement as a scientific process by which we understand the world, the thin conception of validity looks like throwing the baby out with the bathwater. For example, Borsboom and Markus (2013) worry that, insofar as measurement is supposed to generate knowledge (as justified true belief) about the world, giving up on the truth condition means losing our grasp on reality.

However, if we step away from the battlefield for a moment, we can see that the point of contention does not run as deep as the heat of battle might suggest. The thick camp has never denied the value of validation tailored to specific use contexts, just as the thin camp has never refused a truth claim when it’s on offer. Some (Hood 2009; Cizek 2012) have suggested ways of accommodating both conceptions, where validity in terms of capturing true constructs sits at the core of validity theory and goal-specific validations help bridge theory with use contexts. Some theories of causation routinely invoke contextual information (e.g., Woodward 2007), and Larroulet Philippi (2021) goes as far as arguing that it is impossible to have a

context-independent causal account of validity at all. In other words, there is nothing intrinsically inconsistent for a theory to both provide causal information and do so in a context-sensitive way. The ideal measurement should be both true and useful.

The dispute, instead, is about a much more surface-level problem that, at the same time, has a much greater potential to cause harm. Shepard (2016) points to an important stake in this fight that provides a positive reason for giving up validity's implication for truth:

Having taught policy-makers, citizens and the courts to use the word validity, especially in high-stakes applications, we cannot after the fact substitute a more limited, technical definition of validity.

She goes on to cite several legal cases that rely on an understanding of validity as goal specific, such as the 1971 *Griggs v. Duke Power Co.* case, where the US Supreme Court ruled against the use of intelligence tests to select employees for higher-level jobs because, as Justice Burger explained, tests were supposed to evaluate people only in their capacity as employees, not people in any holistic sense. Shepard argues that because tests are routinely evaluated in the thin, argument-based sense, it is epistemically irresponsible to sneak truth in retrospectively.

Recall that the IBE argument from validity to realism relies on the assumption that measurement is a descriptive project that aims at capturing some target that objectively exists in the world. To call a measurement valid is to claim that one has succeeded in this task, which implies the existence of this target that is supposed to have been successfully captured. As I have argued in section 2, the view that measurement is a descriptive project about some objectively existing part of the world does not fit many measurement situations. Consequently, to say that success in measurement implies a particular way the world has to be is both epistemically irresponsible and, as Shepard points out, politically dangerous.

Nevertheless, I don't think we should see this as a simple defeat in our theorizing about measurement. The correspondence framework fails to provide a fruitful account of how measurement works not only because measurement often fails to correspond but also because measurement often succeeds in doing a lot more than corresponding. What we need is to develop theories of measurement that properly respect its creative powers. In what follows, I turn to my positive thesis. I argue that, far from a mere concession in the face of practical challenges, the thin conception of validity can serve as the basis of a new, and hopefully more scientifically fruitful, kind of measurement theory.

Although the construct-validity program is most frequently associated with the realist ontology I have been resisting, let us take a moment to remember its logical positivist roots. In Cronbach and Meehl's original conception, the construct is a node in a nomological network of other constructs, all of which eventually trace to some verificationist, empiricist interface with the world. In other words, the meaning of a construct is exhausted by its observational consequences. Whether we should make an ontological commitment to a construct depends on whether positing the construct is scientifically useful, given our observations in the world and the theory's predictive and explanatory powers. The goal is not to find constructs that correspond with entities in the world; that would be doing metaphysics, after all.

Because the meaning of a construct is given by its theory, this view has the standard problem that plagues any project with a verificationist sense of meaning—validation of tests is always internal to the theory. “A consumer of the test who rejects the author’s theory cannot accept the author’s validation,” explains Cronbach and Meehl (1955). Nevertheless, we may salvage one essential attitude of this program: that the ontology of a construct must be built on how useful it is in a broader scientific system.

Instead of taking the existence of the construct as the precondition for measurement and validity, I propose a reversal of the inferential order. According to the argument-based theory of validity, a test is valid if it is useful in the right contexts. Because tests are almost always developed with specific uses in mind, there should be little ambiguity in determining the validity of a test.

In many testing situations, it is enough to know that a test is valid for its purpose. Sometimes, however, we may want to develop theories for the purpose of adapting test scores across context or offering unified explanations of multiple valid tests. These measurement theories can be developed by reflecting upon the tests’ design principles. They may posit constructs, causes, or any other theoretical entities often employed to explain and unify phenomena. Their qualities are then judged by how they cohere with phenomena in the usual way, with “phenomena” in this case being valid (useful) results. If a theory is sufficiently scientifically powerful, the reality of its posits can be discussed in the same way we assess the reality of other nonobservables.

Put in a different way, the correspondence framework of measurement is a foundationalist framework with the target of measurement at the foundation. The reality of the target grounds its measurability conditions, which give rise to a measurement theory. The measurement theory posits a particular kind of relationship between numbers and the measured reality, which defines a validity theory. In contrast to this picture, my proposed validity-first framework puts validity judgments at the foundation. Validity judgments, understood in the thin, argument-based way, are practically easy to obtain but epistemically insubstantial. A measurement theory—multiple measurement theories—can be built on top of validity claims through the unification and theorization of valid tests. Finally, if some of these measurement theories are sufficiently scientifically powerful, we may use them to justify existential claims about the unobserved constructs that play explanatory roles in these theories.

This new, “validity-first” framework of measurement has several advantages over the correspondence framework. First, in the validity-first framework, the theoretical commitment increases with the evidential burden. Instead of starting with an assumption that the world is a particular way, we start by answering a small, well-defined question (Does this test do what its developers want it to do?), the answer to which provides a small piece of the puzzle (that this test is valid in this particular context). To make contentious claims, such as that a construct exists in the world and admits a total order, we would need not only a lot of empirical evidence about valid tests but also a lot of theorizing. In other words, it is easier to be certain of the usefulness of a test than it is to be certain about the objective existence of a construct.

Second, the validity-first framework does not depend on any particular view of reality. In the correspondence framework of measurement, measurement is made possible by the assumption that the world bears some kind of relationship with

the measurement results, which depends on the world being a certain way that allows for this kind of relationship. For example, in the RTM, a construct is representable by numbers if it can be axiomatized in a certain way. In IRT, a construct is measured by a test if it is causally responsible for the test results. By contrast, the validity-first framework is agnostic about whether a valid test is a measurement at all until we construct a theory about why it would be fruitful to consider these test results as measurement. The theory will then have to be evaluated before it is accepted. If we have a specific commitment about the nature of truth or the structure of the world, it will be reflected in the theories we propose.

Finally, the validity-first framework is more descriptively apt for real measurement situations. As discussed in section 2, measurement interacts with the world in complex ways. An accurate description of the world is often not the main driving force behind the development, implementation, and assessment of a test. Sometimes, trying to measure the world in a certain way can profoundly alter the world in the process. The scientific reality requires us to have a more flexible view of what is achieved when a piece of measurement is deemed successful. The validity-first framework offers that flexibility by giving us space to theorize about how a piece of measurement is successful.

#### 4. Conclusion

We often see measurement as a kind of mediated perception aimed at providing information about a part of the world. Because philosophers of science have grown accustomed to dealing with problems affecting mediated perceptions, it is tempting to discuss measurement in the same terms. Compared with more direct forms of perception, measurement allows for a greater risk of theory-ladenness, is more susceptible to inductive failures, and presents a greater challenge for noncircular verification of results, among other issues. These problems have, by and large, dictated past theorizations about measurement. For example, the operationalist's answer to the inductive failure of an instrument is to define constructs by their forms of measurement, so the very failure itself means that it's not that the measurement fails to work; it's that the target of measurement has changed.

This way of theorizing about measurement gives it both too much and too little credit. It gives measurement too much credit by assuming that it is a straightforward, albeit lossy, way of describing nature. It assumes that the process of measurement is a well-defined, self-contained scientific process—that we are always sure which part of the world we are describing and what the descriptive process involves before we start. As I have argued throughout this article, these assumptions are often mistaken.

At the same time, seeing measurement as mediated perception gives measurement too little credit by ignoring its ability to profoundly shape both our theorizing about the world and the world itself. The amount of creativity that often goes into a piece of measurement is dismissed as theoretically uninteresting—if the results can only be “correct” or “incorrect,” as descriptive projects often are, then the need to exercise creativity is largely a weakness, not a strength.

As I have argued in this article, it is more fruitful to operate without a preconceived measurement theory. Because test development is often driven by practical concerns, a successful theory must respect the impact of practice. This is already



reflected in the testing community's turn toward the more practically focused argument-based approach to validation. By making validation the foundation of measurement theory, we can better make sure that whatever epistemic or metaphysical conclusions we draw from measurement are properly grounded.

In conclusion, this article examines the role existence plays in measurement validity. I reviewed existing popular theories of measurement and validity, and I argued that they all follow a correspondence framework, which starts by assuming that an entity exists in the real world and has certain properties that allow it to be measurable. To measure is to passively document the entity's properties, and to measure validly is for this documentation to be accurate. By looking at debates from within the testing community and drawing on literature from the sociology of measurement, I showed that the correspondence framework faces both a theoretical challenge, where the assumption of the existence of the entity is rarely justifiable, and a practical challenge, where it does not match how measurement is done in many high-stakes situations. In its place, I suggested the validity-first framework of measurement, which reverses the justificatory order. I argued that we ought to start with a practice-based validation process, which serves as the basis for a measurement theory and only posits objective existence when it is scientifically useful to do so.

**Acknowledgments.** This manuscript was completed during the COVID-19 pandemic. I am extremely grateful for the patience on behalf of the editors and reviewers as I struggled through this process.

I'd like to thank Michael Schneider, Greg Lauro, Simon Huttegger, Cailin O'Connor, and Chris Mitsch for invaluable discussions. I benefited greatly from reading group discussions with Adam Chin, True Gibson, David Mwakima, and Jingyi Wu. I'd like to also thank Chrysi Xiyu Du and an audience at the 2022 Social Science Roundtable for their support.

## References

- Alexandrova, Anna, and Daniel M. Haybron. 2016. "Is Construct Validation Valid?" *Philosophy of Science* 83 (5):1098–109.
- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- American Psychological Association. 1954. "Technical Recommendations for Psychological Tests and Diagnostic Techniques." *Psychological Bulletin* 51 (2):1–38.
- Anastasi, Anne. 1950. "The Concept of Validity in the Interpretation of Test Scores." *Educational and Psychological Measurement* 10 (1):67–78.
- Baccelli, Jean. 2020. "Beyond the Metrological Viewpoint." *Studies in History and Philosophy of Science Part A* 80:56–61.
- Boring, Edwin G. 1923. "Intelligence as the Tests Test It." *New Republic* 35 (6):35–37.
- Borsboom, Denny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. New York: Cambridge University Press.
- Borsboom, Denny, and Keith A. Markus. 2013. "Truth and Evidence in Validity Theory." *Journal of Educational Measurement* 50 (1):110–14.
- Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111 (4):1061–71.
- Boyle, Gregory J. 2008. "Critique of the Five-Factor Model of Personality." In *The SAGE Handbook of Personality Theory and Assessment*, edited by Gregory J. Boyle, Gerald Matthews, and Donald H. Saklofske, 295–312. Los Angeles, CA: SAGE Publications.
- Bradburn, Norman M., Nancy Cartwright, and Jonathan Fuller. 2017. "A Theory of Measurement." In *Measurement in Medicine: Philosophical Essays on Assessment and Evaluation*, edited by Leah McClimans, 73–88. London: Rowman & Littlefield International.

- Brennan, Robert L., editor. 2006. *Educational Measurement*. 4th ed. Westport, CT: Praeger.
- Campbell, Norman R., and Harold Jeffreys. 1938. "Symposium: Measurement and Its Importance for Philosophy I." *Proceedings of the Aristotelian Society, Supplementary Volumes* 17:121–51.
- Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.
- Cizek, Gregory J. 2012. "Defining and Distinguishing Validity: Interpretations of Score Meaning and Justifications of Test Use." *Psychological Methods* 17 (1):31–43.
- Costa, Paul T. Jr., and Robert R. McCrae. 1992. "Four Ways Five Factors Are Basic." *Personality and Individual Differences* 13 (6):653–65.
- Cronbach, Lee J. 1980. "Validity on Parole: How Can We Go Straight?" 99–108. San Francisco: Jossey-Bass.
- Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4):281–302.
- Eysenck, Hans Jurgen. 1991. "Dimensions of Personality: 16, 5 or 3? Criteria for a Taxonomic Paradigm." *Personality and Individual Differences* 12 (8):773–90.
- Guilford, Joy P. 1946. "New Standards for Test Evaluation." *Educational and Psychological Measurement* 6 (4):427–38.
- Hardcastle, Gary L. 1995. "S. S. Stevens and the Origins of Operationism." *Philosophy of Science* 62 (3): 404–24.
- Hood, S. Brian. 2009. "Validity in Psychological Testing and Scientific Realism." *Theory & Psychology* 19 (4):451–73.
- John, Stephen. 2022. "Why Five Fruit and Veg a Day? Communicating, Deceiving, and Manipulating with Numbers." In *Limits of the Numerical: The Abuses and Uses of Quantification*, edited by Christopher Newfield, Anna Alexandrova, and Stephen John, 141–60. Chicago: University of Chicago Press.
- Kane, Michael T. 2013a. "Validating the Interpretations and Uses of Test Scores." *Journal of Educational Measurement* 50 (1):1–73.
- Kane, Michael T. 2013b. "Validation as a Pragmatic, Scientific Activity." *Journal of Educational Measurement* 50 (1):115–22.
- Larroulet Philippi, Cristian. 2021. "Valid for What? On the Very Idea of Unconditional Validity." *Philosophy of the Social Sciences* 51 (2):151–75.
- Loevinger, Jane. 1957. "Objective Tests as Instruments of Psychological Theory." *Psychological Reports* 3 (3):635–94.
- McCrae, Robert R., and Oliver P. John. 1992. "An Introduction to the Five-Factor Model and Its Applications." *Journal of Personality* 60 (2):175–215.
- Merry, Sally Engle. 2016. *The Seductions of Quantification: Measuring Human Rights, Gender Violence, and Sex Trafficking*. Chicago: University of Chicago Press.
- Messick, Samuel. 1989. "Validity." In *Educational Measurement*, edited by Robert L. Linn, 13–103. New York: American Council on Education and Macmillan.
- Micell, Joel. 1999. *Measurement in Psychology: A Critical History of a Methodological Concept*. Vol. 53. Cambridge: Cambridge University Press.
- Mickelson, Kristin D., Ronald C. Kessler, and Phillip R. Shaver. 1997. "Adult Attachment in a Nationally Representative Sample." *Journal of Personality and Social Psychology* 73 (5):1092–106.
- Porter, Theodore M. 1996. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton, NJ: Princeton University Press.
- Scott, James C. 1996. "State Simplifications: Nature, Space, and People." *Nomos* 38:42–85.
- Shepard, Lorrie A. 1997. "The Centrality of Test Use and Consequences for Test Validity." *Educational Measurement: Issues and Practice* 16 (2):5–24.
- Shepard, Lorrie A. 2016. "Evaluating Test Validity: Reprise and Progress." *Assessment in Education: Principles, Policy & Practice* 23 (2):268–80.
- Siegel, Reva B. 1994. "Home as Work: The First Woman's Rights Claims Concerning Wives' Household Labor, 1850–1880." *The Yale Law Journal* 103 (5):1073–217.
- Spearman, Charles E. 1904. "'General Intelligence,' Objectively Determined and Measured." *American Journal of Psychology* 15 (2):201–92.
- Stevens, Stanley Smith. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684):677–80.
- Stevens, Stanley Smith. 1968. "Measurement, Statistics, and the Schemapiric View." *Science* 161 (3844):849–56.

- Stone, Caroline. 2019. "A Defense and Definition of Construct Validity in Psychology." *Philosophy of Science* 86 (5):1250–61.
- Suppes, Patrick, and Joseph L. Zinnes. 1963. "Basic Measurement Theory." In *Handbook of Mathematical Psychology*, Vol. I, edited by Robert Duncan Luce, Robert R. Bush, and Eugene Galanter, 1–76. New York and London: John Wiley and Sons.
- Tal, Eran. 2019. "Individuating Quantities." *Philosophical Studies* 176 (4):853–78.
- Vessonen, Elina. 2021. "Representation in Measurement." *European Journal for Philosophy of Science* 11 (3):1–23.
- von Helmholtz, Hermann. (1887) 1930. *Counting and Measuring*. Translated by Charlotte Lowe Bryan. New York: D. Van Nostrand Company.
- Woodward, James. 2007. "Causation with a Human Face." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by Huw Price and Richard Corry, 66–105. Oxford: Oxford University Press.