Article

Philosophy of the Social Sciences
2021, Vol. 51(2) 151–175

DOI: 10.1177/0048393120971169
journals.sagepub.com/home/pos


Valid for What? On the Very Idea of Unconditional Validity

Cristian Larroulet Philippi[1]

Abstract

What is a *valid* measuring instrument? Recent philosophy has attended to logic of justification of measures, such as construct validation, but not to the question of what it means for an instrument to be a valid measure of a construct. A prominent approach grounds validity in the existence of a causal link between the attribute and its detectable manifestations. Some of its proponents claim that, therefore, validity does not depend on pragmatics and research context. In this paper, I cast doubt on the possibility of a context-independent causal account of validity (what I call unconditional validity). I assess several versions, arguing that all of them fail to judge the validity of measuring instruments correctly. Because different research purposes require different properties from measuring instruments, no account of validity succeeds without referring to the specific research purpose that creates the need for measurement in the first place.

Keywords

Measurement, Construct Validity, Causation, Social Sciences, Psychometrics

1. Introduction

What is a valid measurement? When are measuring instruments or procedures valid? In everyday research, 'valid' connotes the approval of a measuring

Received 10 October 2020

[1]University of Cambridge, Cambridge, UK

Corresponding Author:
Cristian Larroulet Philippi, Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK.
Email: cristianlarroulet@gmail.com

instrument. For most practitioners, the intuitive idea behind the notion of validity consists in whether an instrument actually measures the construct at stake (originally stated in Kelley 1927). This intuitive idea has been made precise in quite distinct ways by different authors, and there is no settled consensus on a preferred account. Now, whether a particular measuring instrument—for example, a happiness questionnaire—is considered valid/invalid has not only research consequences, but also policy implications (Angner 2011b). Thus, validity's definition matters.

Enter causation. For years considered a notion difficult to pin down, it is nowadays not uncommon to use causation as a primitive to define other concepts—such as "scientific explanation," "events" (Davidson 1969), or "verisimilitude" (Northcott 2013). Causation—and its close kin correlation—have also been used to define validity. Most famously, psychometrician Denny Borsboom and colleagues (Borsboom 2005; Borsboom et al. 2004 [henceforth, BMH]) stated that a measuring procedure "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" (BMH, 1061). In turn, philosopher Matthias Michel (2019) defines validity using a correlational condition. This usage of causation (or correlation) has prima facie appeal. Consider a mercury thermometer. Plausibly, the fact that higher (lower) temperatures cause higher (lower) levels of the volume of mercury is what makes a mercury thermometer valid for measuring temperature.

Part of Borsboom's motivation to define validity using causation is to keep validity invariant to the pragmatic and inferential aspects of the research context within which measurements are performed. The intended contrast here are the leading accounts of validity in educational research (e.g., Kane 2006; Messick 1989). These accounts define validity in ways that include the possible interpretations and usages to which the measurement results would be put. To Borsboom, these accounts conflate validity with epistemic and pragmatic considerations, and leave the theorist and the practitioner with a daunting job and without clear guidance (BMH, 1061). As BMH put it, "validity is not complex, faceted, or dependent on . . .[the] social consequences of testing . . . [but] about the simple, factual question of whether [an instrument] measures an attribute" (1061). Causation is BMH's bet for providing a definition of validity that satisfies this description.

Let us call causal accounts of validity that are invariant to context-specific aspects, such as BMH's, accounts of *unconditional validity*. At first sight, validity might seem to be a property that measuring instruments either have or not independently of context. Thus, articulating an adequate account of

unconditional validity should not be difficult. However, I will argue that the need to include context into validity's definition remains even within a causal framework. Especially relevant about the research context is what I call here the "research purpose": the inferences and/or actions that create the need for measurement in any given case. I argue that the need to include the research context goes deeper than in other well-known cases, such as probabilistic causation, where judgments of causation are relativized to specific populations, and pragmatic encroachment accounts of knowledge, where judgments of justified belief are relativized to the contextual stakes (determined by agents' purposes). It's not only the case that the validity of an instrument depends on the population and circumstances in which it is used, and that whether an instrument is "valid enough" depends on how demanding the research purposes in place are. If we think of validity in graded terms, even the *degree* of validity of a measuring instrument can depend on those purposes. Because different research purposes require different properties from a measuring instrument, we sometimes cannot judge instruments' degrees of validity without referring to the specific research purpose that creates the need for measurement in the first place.

I start in section 2 clarifying our object of study by distinguishing it from validation. Then, in section 3, I narrow down the space of possibilities regarding causal accounts of validity. In sections 4 and 5 I introduce and reject two accounts of validity that are arguably in the market—each being at opposite extremes of the causal spectrum here considered. If neither extreme is good, perhaps something in between might do the job. I show in section 6 that no *unconditional* account of validity succeeds.

2. Validity versus Validation

We need to distinguish validity from validation. This is important for two reasons. Many definitions of validity given by psychometricians and philosophers are, arguably, definitions of validation. Moreover, validation has received most of the philosophical attention.

Roughly speaking, the validity/validation distinction maps to the *true* belief versus *justified* belief distinction. Recall the intuitive gloss given to validity, namely, whether an instrument measures the construct it intends to measure. Nothing there suggests that validity is an epistemic term. One thing is to say that an instrument *actually measures* an attribute; quite another that we are *justified in believing* so. As Markus and Borsboom remark, validity should be understood "in terms of what holds true independent of the available evidence rather than as a summary of the extent to which evidence supports the belief" (2013, 14).

Yet this has not been always so. Either due to positivistic influences (Cronbach and Meehl 1955; see discussion in Borsboom 2005; Markus and Borsboom 2013) or just to avoid making ontological claims, several theorists have preferred defining validity in epistemic terms. A clear example is Messick's account.[1] "Validity," says Messick,

> is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on [measurement results]. As such, validity is an inductive summary of both the existing evidence for and the potential consequences of [measurement results'] interpretation and use. (1989, 13)

Messick's "validity" speaks about our epistemic justification for measurement-based inferences and actions. Thus his definition is better conceived as an (expansive) account of *validation*. Philosophers have also failed to distinguish validity from validation. For example, Stone's (2019, 1253) proposed account of construct validity says: "A measure, *m*, of some construct, *C*, is construct valid if and only if we are sufficiently justified in believing *m* tracks *C*." Evidently, Stone's "construct validity" is also an epistemic concept. Thus her account is also better conceived as an account of validation.[2]

It should now be clear that recent philosophical discussion under the banner of 'construct validity' has largely been on validation (an exception: Hood 2009). Think of Angner's (2011a) exposition of the psychometric approach to measurement, Alexandrova and Haybron's (2016) coherentist defense of "construct validity," and Vessonen's (2017) discussion of "procedural validity." These are all discussions chiefly about the epistemic project that construct validation consists in.

I shall stress that definitions of validity are not necessarily without epistemological import. They shape the epistemic task of validation. To see how, recall Stone's (2019) definition. Arguably, it tacitly defines (the ontological sense of) validity as "tracking": an instrument measures the intended construct iff it tracks it. Now, Stone doesn't explicate what "tracks" amount to. As shown below, however, this *does* merit attention. The way in which "track" is precisified—more demandingly or less so—will affect when are we justified in believing that the measurement "tracks" the attribute.

---

[1]See Hood (2009) for further examples and discussion.

[2]Stone's and Messick's accounts differ. Stone's definition focuses only in believing whether *m* tracks *C*. Thus it excludes what Messick's account includes: beliefs related to researchers' inferential purposes (e.g., did academic achievement improved over the year?) and pragmatic purposes (e.g., whether this particular student should pass).

To be clear, my focus here is not epistemic—I don't assess construct validation practices. I've only clarified that accounts of validity matter for validation practices.

3. Validity and Causation

We can now narrow our target. First, I explicitly discuss *causal* accounts, not correlational ones. Recall BMH's condition (b): the attribute has to causally produce variations in the outcomes of the measurement procedure. This excludes from candidates for validity so-called "indirect" (or "proxy") instruments (Angner 2011b, 128).[3] I follow BMH here to simplify the presentation. The core issue discussed below is the strength of the relation between attribute and measurement outcomes. Since this issue applies equally to causal and correlational definitions, my arguments below would straightforwardly apply to correlational versions of the causal definitions considered. Whether we should distinguish between "indirect" and "direct" measurements is not germane to our discussion.

Second, providing unambiguous causal accounts requires making the term 'cause' (simpliciter) precise. This requires fixing its scope. Typically, variables have causal effects in some background contexts but not in others. Does C cause E simpliciter when C causes E in *all* background contexts? In *most* background contexts? Or is *one* background context enough? The need to relativize causal claims to background contexts should be familiar from the probabilistic causality literature (e.g., Fitelson and Hitchcock 2011, 601). Imagine quite realistically that an attribute (*X*) causally affects certain measurement outcomes (*M*) in some but not all background contexts (e.g., for a given population in a given situation).[4] Does BMH's condition (b) hold in such cases? As stated, BMH's account doesn't tell us.

---

[3]Most philosophers do not privilege causation over correlation. For example, Angner (2011b) and Alexandrova (2017) consider (different) measuring instruments of happiness as potentially valid for measuring well-being despite their measurement outcomes not being *effects* of well-being. (These outcomes, these authors say, share a common cause with well-being.) Likewise, Cartwright and Bradburn (2011) treat measurement outcomes that share a common cause with the attribute on a par with those that are effects of the attribute. Compare psychometricians' usual distinction between "formative" and "reflective" measurement models (e.g., BMH 1069).

[4]Henceforth, I always call '*X'* the attribute of interest and '*M'* the outcomes provided by the measuring procedure whose validity is at stake. Also, I use 'measuring instruments' and 'measuring procedures' interchangeably. Presumably, it is the measuring *procedure* (which includes the instrument used) the object of validity assessments. But it is common among authors to speak of, say, "the validity of tests" referring to the validity of a standardized procedure that applies those tests.

However, we shouldn't read (b), or alternative causal conditions, in a wide-scope way. Temperature causally affects the readings of a mercury thermometer within a temperature range—not below the point where mercury freezes. Should we say 'the readings are not causally affected simpliciter,' and thus judge the instrument invalid? A wide-scope view seems plainly unattractive. It's also at odds with research practice. Social science measuring instruments—for example, self-esteem scales, anxiety scales, etc.—that have been developed and validated for some population (e.g., a country), are usually not assumed valid for another population by default. And this is not only an issue of translation: Researchers genuinely ask whether a scale validated in, say, France, is valid (after due translation) in Chile. Thus, instruments may be judged valid for a given population *even if* they may not for another.[5]

So, here is a first way in which validity judgments must be restricted to context-specific situations: to a given data-background context. From now on, then, when we discuss whether a measuring instrument is valid, we do so assuming a given data-background context.[6] Accordingly, that mercury thermometers' readings are not causally affected by temperature below mercury's freezing point is not necessarily inconsistent with saying 'these thermometers validly measure temperature'—we need to disambiguate that judgment in terms of the data-background context assumed before we can assess it. Likewise, the fact that the scores of a test are not causally affected by intelligence in, say, the case of respondents that aren't able to read is not necessarily inconsistent with saying 'this test validly measures intelligence.' The interesting question is whether there are *further* contextual aspects, beyond the data-background, that must be part of an account of validity.

Finally, I set one issue aside. Recall that BMH's account had two conditions; condition (a) explicitly requires the existence of the attribute in question. Making sense of measurement practices, say BMH (1063), requires this commitment to ontological realism. But this point remains controversial in the philosophical literature. Here, I do not enter into this debate.[7] Throughout

[5]See Alexandrova (2017, 125). Of course, I'm assuming that at least some attributes studied by scientists are not defined in a population-specific way (e.g., French-anxiety).

[6]This move is standard in discussions of causation. See, e.g., Fitelson and Hitchcock (2011, 601), Northcott (2013, 1472-3), and Sprenger (2018, 376-7).

[7]As Hood (2009) notes, strictly speaking, BMH's account entails ontological realism only in conjunction with the claim that valid measurement instruments exist. Still, some authors commenting on Borsboom's program have expressed worries about realist commitments being necessary for making sense of measurement practice (Alexandrova 2017, 148). I concur. For one, certain kinds of anti-realisms (e.g., van Fraassen's) are arguably compatible with causal explanations (see Hitchcock 1992),

the paper, I assume for the sake of argument that the attribute exists, and center exclusively on the causal condition.

4. A Maximalist Account

We can now consider causal accounts with some definiteness. Any such account will be a precisified version of BMH's condition (b), which states that variations in the attribute (*X*) cause variations in the outcomes of the measurement procedure (*M*). Different accounts correspond to different ways of disambiguating the relevance or strength of the causal relation posited. To achieve validity, for instance, do *all* and *only* variations in *X* need to cause variations in *M*? Or is it enough that *X*-variations cause *some M*-variations? BMH's condition (b) doesn't tell us. We will consider, in this and the next section, two possible disambiguations. I call *maximalist* the account which specifies (b) as:

$$\left(\mathrm{b}^{\mathrm{Max}}\right)\text{: } \textit{all} \text{ and } \textit{only } X\text{-variations cause } M\text{-variations}.$$

Condition (b^{Max}) rules out instances where *X*-variations don't lead to *M*-variations (the "*all"* part) and instances where *M*-variations aren't caused by *X*-variations (the "*only"* part). An account with (b^{Max}) might seem appealing. Take again the case of a mercury thermometer. Within some temperature range, the fact that higher (lower) temperature always lead to increases (decreases) in the volume of mercury is crucial for the thermometer's validity. But that is only part of it. That variations in the volume are not caused by other factors (e.g., atmospheric pressure) also seems key. That is, a mercury thermometer's validity seems to depend on it being causally affected by, but only by, temperature.

Alas, (b^{Max}) is too demanding. Think of an educational test of *X*=intelligence. According to (b^{Max}), this test's scores (*M*) need to be affected by *any* variation in *X* and *only* by variations in *X.* Imagine the best intelligence test we might currently (or plausibly) have.[8] Arguably, such a test should come out as valid in any reasonable account of validity. However, does *all* variation in *X* produce variation in *M?* Surely the test has a finite

---

and thus prima facie with causal accounts of validity. Furthermore, Tal (2017) rightly warns about muddling measurement realism with entity realism. See nuanced discussions about realism and psychometrics in (Hood 2013; Vessonen 2019).

[8]Recall that I'm only assuming for argument's sake the existence of the attribute (intelligence). I use the intelligence example because it is prevalent in the literature.

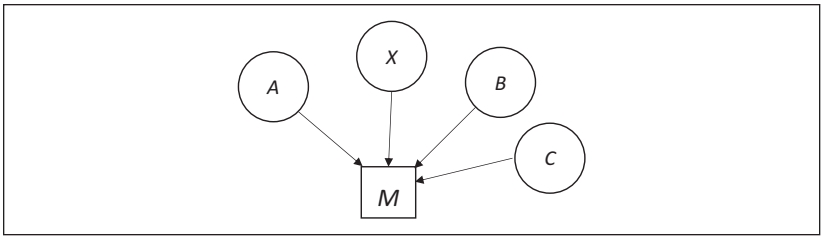

Figure 1. A causal graph of test scores *M*.

number of questions, which entails that, as long as *X* is a continuous variable (or an ordinal one but with a large number of intermediate values), the test will not be able to discriminate *all* possible variations in *X.* This entails a failure of the "*all"* part.[9] Moreover, this demandingness problem also holds for dichotomic attributes. Any medical test that is less than 100% sensitive to the presence of a disease fails the "*all*" part.

What about the "*only*" part? As it is widely recognized by psychometricians in general (e.g., Borsboom and Mellenbergh 2004), measurement in the human sciences inevitably deals with error. The nature of such an error is not a simple issue, but one plausible interpretation is this. Even in our best intelligence tests, we might expect something like the following causal structure. The variables X = intelligence, A = motivation, B = ability to focus on tests, and C = emotional mood jointly determine test scores M (Figure 1).

Researchers might treat *A, B*, and *C* as the source of statistical "error" (or "noise"), especially when they have grounds for thinking that *A, B*, and *C* are not correlated with *X.* In such cases, *M* might be said to capture *X* "unbiasedly," because there is no confounding. But even in this (highly favorable and unlikely) kind of case, we still have *M* being not *only* affected by *X—A, B*, and *C* (the "statistical noise") also affect *M*. Thus, even in this favorable case the "*only"* part fails.

Now, these points surely generalize beyond intelligence tests. Condition (b^{Max}) is just too strict. It's not clear that all of our best instruments even in the physical sciences live up to this condition. If we want an account validity that makes sense of research practice, we cannot ask for such strictness.

Notably, however, Michel (2019) gets very close to endorse a correlational version of (b^{Max}). He states:

[9]Failures of the "*all"* part also happen when some *X*-variations are out of the range in which M is affected, generating floor effects and/or ceiling effects.

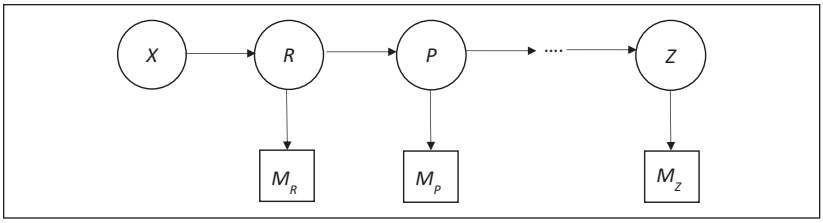

Figure 2. A causal chain of attributes with their respective measurements.

> [A] procedure for measuring consciousness is valid only if variations in indications produced by that procedure [*M*] correlate with variations in the presence or absence of consciousness, or with levels of consciousness [*X*], and do not correlate systematically with other properties that are independent from consciousness. (2019, 1243)

By requiring that *M*-variations do not correlate with other properties, Michel endorses the "*only*" part of (b^{Max}). He does not explicitly sanction the "*all*" part—it's an open question whether being sensitive to only some of *X*-variations makes an instrument invalid for Michel. Regardless, requiring the "*only*" part makes Michel's account too strong to be accepted.

5. A Minimalist Account

At the opposite extreme lies a *minimalist* approach to validity, specifying (b) as follows.

$$\left(b^{Min}\right)\text{: some } X\text{-variations cause some of the } M\text{-variations.}$$

This is the most plausible interpretation of the account that BMH aimed to offer (see subsection *5.1*). However, this account gives judgments about validity that are hard to square with reasonable research practice. Consider, firstly, the causal structure represented in Figure 2, where the variable of interest, *X*, causes *R*, which causes *P*, which . . . (several more variables) . . .. causes *Z*. Since causation is typically transitive (Paul and Hall 2013), it follows that *X* causes *R, X* causes *P*, . . . and *X* causes *Z.* Now, imagine also that for each one of *R, P*, . . . *Z*, we have a valid—under (b^{Min})—measurement. That is, variable *R* causally affects measurement outcomes M_R, *P* affects M_P, etc. By transitivity, *X* causes M_R, *X* causes M_P, . . . and *X* causes M_Z. Since each *M* is causally affected by *X*, each *M* satisfies the condition for validity. Thus, under (b^{Min}), any valid measurement of *R, P.* . .and *Z* is a valid

measurement of X. More generally, *any* valid measurement of a variable that is an effect of *X* is a valid measurement of *X*.

Under (b^{Min}), measurements of variables further down the chain come out as valid *no matter how far down the causal chain they are*. This runs afoul of research practice. Let's say that *X* = intelligence and *Z* = annual income. It seems plausible that in some contexts (say, capitalistic economies with strong public educational systems, strict antidiscrimination employment laws, etc.) variations in intelligence cause some (perhaps small) variations in annual income, through a somewhat long causal chain involving several supporting/mediating factors. Annual income surely affects other variables further down the causal chain, say, leisure travels. (b^{Min}) implies that *any* survey that is a valid measure of leisure travels is a valid measure of intelligence. But we don't see psychologists endorsing instruments such as leisure travels surveys as valid instruments of intelligence. More generally, researchers don't seem to be indifferent to *how far down* the causal chain the measuring instruments are. An account that defines validity with (b^{Min}), however, cannot make sense of this arguably sound (see below) research practice.

Secondly, consider again the causal structure represented in Figure 1. Under (b^{Min}), the test is as much a valid measurement of *X* as it is of *A, B*, and *C*. These judgments are counterintuitive—it is not common research practice to consider a test as a valid measurement of *all* those variables at the same time. But, under (b^{Min}), there are no grounds for singling out intelligence as the only attribute being validly measured. Moreover, test scores are causally affected by many other factors—the caloric content of students' lunches (Figlio and Winicki 2005), air pollution (Ebenstein et al. 2016), and temperature (Graff Zivin et al. 2018) on testing days, etc. Do we see practitioners considering these tests as valid measures of *all* those other variables? Not really. More importantly, in cases where *many* attributes other-than-intelligence are affecting the scores, researchers might be doubtful whether intelligence is in fact being validly measured by test scores. But again (b^{Min}) cannot vindicate such doubts.

Yet these doubts are well-motivated. Researchers should care about *how many* variables other-than-the-attribute-of-interest are causally affecting *M*, if only because they should care about *how much* of the *M*-variations comes from non-*X*-variations. When other variables are behind much of *M*-variations, the measuring instrument tells us little about the attribute of interest. These considerations arguably also explain researchers' lack of indifference to *how far down* the causal chain measuring instruments are. The farther down they are, the more the number of other factors causally affecting them, which (ceteris paribus) leads to a larger share of *M*-variations generated by non-*X*-variations. All said, an account using (b^{Min}) is compatible with judging

measuring instruments that are poor indicators of the attribute of interest as valid. In being so lenient, causal condition (b^{Min}) fails to make sense of sound research practice.

5.1. A Possible Motivation for Minimalism

BMH have a reply. Theirs deflationary account intends to be a *revisionary* project in the following sense. The notion of a valid measuring instrument, they claim, has been confused with that of an "optimal" one. (BMH use "optimal" neither in the sense of "good enough" nor in the sense of "best," but in the sense of "overall good." I stick to their use.) And their aim is to "decouple" (1070) these two notions. A minimalist account, they could claim, accomplishes this aim. Before appraising this response, let me clarify the relation between validity and optimality, since it matters for our discussion.

Decoupling validity and optimality, I argue, requires "going minimalists." One cannot both endorse an account of validity that requires some substantial level of causal strength *and* insist that under this account valid measuring instruments may be more or less optimal (namely, more or less predictive, reliable, measurement-invariance, etc.). In other words, in causal accounts of validity, causal strength and optimality go hand in hand. Think of the extreme case: (b^{Max}). And consider predictive adequacy first. If *all X*-variations produce *M*-variations and *only X*-variations produce *M*-variations, how can the measuring instrument fail to have predictive adequacy? For instance, if *X*=intelligence, and the "true values" of *X* actually predict with high accuracy SAT scores under specific circumstances, I cannot see how it can be both true that a specific test satisfies condition (b^{Max}) *and* that this test's outcomes are not predictive of SAT scores. Where does the gap come from if all other causes have been ruled out?

The same could be said about other concepts behind the notion of optimality. Reliability is defined in different ways, depending on the preferred psychometric theory (Markus and Borsboom 2013). Consider first a common-sense notion of reliability: same inputs of interest (e.g., same level of intelligence), same measurement outcomes (e.g., same test scores). Hasok Chang calls this "comparability" (2004, 77). Unless one entertains the idea that there is pure chance in social processes (and not only ignorance of all the factors), failures of comparability entail that something *other* than *X* affects *M*. If so, the instrument doesn't satisfy (b^{Max}). Consider now Classical Test Theory's (Gulliksen 1950) definition of reliability: the fraction of observed-score variance in the population of test-takers that is true-score variance.

Again, if those variances differ, there must be other factors behind the observed-score variance, and/or not all true-score variance is being captured by the measurement. Either way, (b^{Max}) doesn't hold.

While (b^{Max}) implies the highest level of optimality (recall, in the sense of overall goodness), (b^{Min}) entails no significant level of optimality. Importantly, moving from (b^{Min}) toward (b^{Max}) by requiring some stronger causal relation leads to higher levels of optimality (at least in several of the most prominent dimensions of optimality, such as predictive adequacy, reliability, measurement-invariance). Thus, a causal account of validity keeps "valid" orthogonal to "optimal" only by going minimalist.

Let's return to BMH's response. In their revisionary project, they propose to decouple validity from optimality. As we saw, this requires going minimalist. Regarding the above examples, BMH would respond that the measurements (e.g., leisure travels as recorded in surveys) are indeed valid under their account, but clearly of poor-quality. This is why they seem counterintuitive and contrary to common research practice—we are confusing validity with optimality. Since they offer a revisionary account, the fact that its judgments don't match with current usage doesn't challenge their account, the reply would go. In short, being BMH-valid is not enough for being a good instrument, but it is a necessary condition.

This is in principle a legitimate response. Sometimes we ask of ourselves not how a theoretical concept is being used but how it *should* be used (Haslanger 2000). And this perspective seems particularly relevant when the concept at stake is expected to guide practice, as it is the case of validity. This makes revisionary proposals (versus conceptual analyses) harder to assess. But at the very least, we should ask what is the point of their revisionary project. The point of reforming a theoretical concept is, presumably, to reveal its deeper significance, the point of having it, or its pay-off for the successful achievement of the cognitive and/or practical purposes involving that theoretical concept. Since decoupling validity from optimality doesn't wear its value on its sleeve, BMH still owe us a reason to prefer their account. Here are two reasons against preferring their account.

In order for there to be a pay-off in decoupling validity from optimality, there has to be some value in singling out instruments that are valid independently of their overall goodness (optimality). What that value might be is not clear to me. I think that the kind of measurements we are considering are largely valued to the extent they enable sound inferences—for example, causal or descriptive inferences about different social groups. But if that is what gives value to our measurements, it is not clear what value could there be in singling out BMH-valid-but-not-optimal measurements from the pool of non-optimal measurements. Neither kind of measurements helps us warranting our inferences of interest.

More importantly, since the notion of validity is used in everyday research to certify measuring instruments, this reform threatens to confuse more than to clarify. To illustrate, as Hood remarks, "[m]any of the objections to mental assessment are charges of invalidity—that the tests are biased in some way, that the inferences made from test scores are unwarranted, or that psychological tests do not measure what they purport to measure" (2009, 451). BMH's reformation would force these critics to say that, 'though the tests are biased, they are still valid,' and that 'test scores are valid despite the inferences from them being unjustified.' Given that validity is commonly used to approve usage of a measuring instrument, this reform invites confusion. All said, I fail to see how the pay-offs this reform offers can surpass the risks of it.

Perhaps readers disagree, and side with BMH to the effect that validity and optimality should be kept orthogonal and that (b^{Min}) provides the best account of validity. If so, they are invited to re-describe what follows as an argument against accounts of unconditional *optimality*. My purpose is to argue that there is no (causally defined) unconditional overall goodness, however we call it. It is not the term that interests me, but the notion that defines overall goodness when it comes to measurement instruments.

6. Something in Between?

Before assessing other accounts, let me propose a test that any account must satisfy. This will make my argument clearer.

6.1. The Point of Validity

Why do we want a concept of measurement validity? What do we want it for? As already hinted at, we need this concept for certifying measuring instruments. That is, we use the term to communicate that they serve our measurement purposes well. I propose, more precisely, that the point of the notion of validity is to certify measurement inferences in the following sense. A valid measurement of X allows us to correctly infer claims about X that are key for the achievement of the goal for which we need a measurement of X in the first place. The research context that leads to a specific measurement endeavor presupposes goals, theoretical (e.g., the testing of scientific hypotheses, the identification of the parameters of a model) and/or policy-oriented (e.g., classifying households as falling below a "poverty line"). These goals determine which claims researchers are interested in, which, in turn, determine what is required of measuring instruments for them to be valid.

To illustrate: if testing a specific hypothesis requires a very precise (say, accurate to the second decimal) measurement of X, any measuring instrument

that is as precise as needed will provide us with measurements that are *valid for* the purpose at hand, even if the instrument is not as precise as it could be (say, accurate to the third decimal). Conversely, no measuring instrument of *X* that is less precise than needed will provide us with measurements that are *valid for* the purpose at hand. In this example, what is required from the measuring instruments is a specific level of accuracy. High levels of accuracy can be relevant in some theoretical contexts (say, for determining physical constants).[10] Other research purposes might imply requirements about the level of the measurement scale (e.g., whether the judgments of interest require measuring the attribute in a ratio scale, or an interval scale, or simply on an ordinal one). Yet others might require different degrees of reliability, or of unbiasedness with respects to different social groups. In these and other ways, the research purpose determines what is required of the measuring instrument so as to provide an outcome that is *valid for* making the pertinent inferences.

The idea that a measuring instrument may be valid for some (less demanding) purposes but not for other (more demanding) purposes, is, I believe, central for this discussion. Recall that I rejected accounts with (b^{Max}) as implausible because most (if not all) of our best measuring instruments of the social sciences will come out as invalid. However, what I'm suggesting here–as a standard of assessment for validity accounts—entails that we should reject accounts with (b^{Max}) *even if* we could develop some measuring instruments that actually satisfy (b^{Max}) (i.e., that are sensitive to *all* and *only* variations in *X*; a big if). This is so because we should be reluctant to classify as invalid instruments that fail to attain perfect sensitivity when they are sensitive enough for the relevant purposes.

Thus, what I'm suggesting is the following. Ultimately, what we want from the concept of validity is to certify measuring instruments for the specific research purpose that creates the need for measurement in the first place. What is of *final value* is not whether this instrument is unconditionally valid—namely, valid for *all* possible purposes we can use measurements for—but whether it is *valid for* the purpose(s) at hand. Now, this suggestion is not an argument against all accounts of unconditional validity. But it does provide a test for any such account: to be successful, such an account has to ground the *valid for* claims that we ultimately care of.[11]

How do the accounts discussed above fare under this test? On one hand, accounts using condition (b^{Max}) fail this test because in being too strict, they are not sensitive to the fact that a measuring instrument may be *valid for*

[10]See Hacking's (1983, Ch. 14) for examples in physics of research contexts that required high levels of measurement accuracy for their theoretical purposes.

[11]This proposal is, of course, sympathetic to Messick's approach to construct validity. However, as argued, Messick's account is about validation, not validity.

some specific purposes while not for other purposes. Such instruments should not come out as invalid across the board, since they are *valid for* some purposes. Accounts using condition (b^{Min}), on the other hand, fail because in being too lenient, they deem valid across the board measuring instruments that in fact are not *valid for* some (even most) purposes. What remains to be seen is whether accounts that are somehow in between, with regards to the strength of the causal condition, can pass our test. We don't need to assess all possible accounts one by one, since there are, to my mind, only two possible kinds of accounts of unconditional validity: categorical and graded. I'll argue that neither can ground the *valid for* judgments across the board.

6.2. Categorical Validity

Both (b^{Min}) and (b^{Max}) are species of a single genus—that of *categorical* accounts of validity, in which measuring instruments are either valid or not. Any categorical account needs to fix a required degree or level of optimality (in terms of causal strength). That required level works as a threshold: instruments that have levels of optimality above the specified level are judged valid. A general statement of categorical accounts' causal condition is:

$\left(b^{\text{Categorical}}\right)$: the instrument's degree of optimality in terms of causal strength $\left(\text{e.g., the degree to which } X\text{-variations cause } M\text{-variations}\right)$ is above a fixed level.

Both (b^{Min}) and (b^{Max}) are categorical accounts. They differ in the level they fix. The former fixes the lowest level, the later fixes the highest level. Would an account that fixes its level somewhere in between succeed as a satisfactory unconditional account of validity?

Where should that required level of optimality be fixed? It seems difficult to justify any univocal answer. Indeed, there is no correct answer to this question—fixing the level anywhere is problematic. This is so because most instruments (currently in use and likely to come) might be valid for some (less demanding) purposes but not for other (more demanding) ones, and because our research purposes vary substantially with regards to the optimality they require. Thus, fixing the required level anywhere would be problematic in a similar sense that using the extreme levels (b^{Max}) and (b^{Min}) was.

For illustration, consider the case of the Hamilton Depression Rating Scale (HAMD). HAMD is composed of seventeen questions related to depression. Most questions are about things that can plausibly be considered effects of depression, such as moods and feelings related to depression,

suicidality, etc. HAMD is widely used in research about depression. Assume for the sake of the argument that depression causally affects (in some substantial sense) the answers to this set of questions.

It is not hard to imagine that there are some purposes for which HAMD is a valid instrument. Take the case where our purpose is to investigate whether one group of people (say, people living in one region) has a higher rate of depression than other group of people (from another region). In line with the standard view on measurement scales, arguably we should not assume HAMD to be a quantitative representation (i.e., interval or ratio scale) of depression. Accordingly, computed averages of HAMD are not, in Stevens' (1946) famous words, "permissible statistics." Nevertheless, this doesn't mean we cannot compare depression (qualitatively) across groups. If HAMD satisfies the conditions for an ordinal scale in the study population, and the distribution of HAMD scores of the two groups satisfy some conditions, then we can validly infer that one group has a higher rate of depression than the other (see Larroulet Philippi ms. for these conditions). That is, HAMD is a measurement instrument *valid for* making that judgment under those conditions. This is a judgment that may be relevant for both theoretical and/or policy-oriented purposes (such as testing hypotheses about depression, or prioritizing mental-health resources).

However, despite having some degree of optimality sufficient for some purposes, HAMD is not valid for other purposes that we would like it to be valid for. In particular, as argued by Stegenga (2018), HAMD is not always a *valid* instrument *for* assessing the effectiveness for targeting depression of alleged anti-depressants that target some of depression's symptoms. One of the effects of depression is insomnia. Accordingly, HAMD includes questions concerning sleeping patterns. However, changes in sleeping patterns (improvements, say) are not *only* affected by depression; they can be directly affected without affecting depression. This means that HAMD scores are sensitive not only to depression, but also to factors that affect sleeping patterns without affecting depression. This is just one corollary of HAMD not satisfying the "*only*" part of condition (b^{Max}), just like most instruments in the social and biomedical sciences. As Stegenga (2018, 116) explains, given this failure of specificity, an alleged anti-depressant that is actually a soporific (i.e., a drug that doesn't affect depression but affects sleeping patterns) might be judged "effective" for treating depression.

Thus, although HAMD is *valid for* some (less demanding) purposes, it is not so for other (more demanding) ones. This is a feature that most—if not all—current measuring instruments of the social sciences share. Measuring instruments don't typically satisfy the "*only"* part of condition (b^{Max}); and the purposes for which we need measurements vary in their demandingness. The

upshot: setting up a *fixed* level of optimality above which an instrument is valid leaves the concept valid useful for only some purposes: those that need no more optimality than that required by the fixed level. Since our research purposes vary substantially with regards to the level of optimality they require, there seems to be no correct answer to the question *Where should we draw the validity threshold?*

The above discussion casts doubt on any account of validity that uses a *fixed* level of optimality for determining whether instruments are (categorically and unconditionally) valid. One obvious response is to have a *context-dependent* threshold: we draw the optimality requirement according to the research purpose. Under such an account, it could still be insisted that an instrument's *degree* of optimality is invariant to the research context. What depends on the research context is only the threshold, that is, the level above which the instrument is considered valid.[12] The causal condition for validity under this account is:

$\left(b^{\text{Contextual}}\right)$: the instrument's degree of optimality in terms of causal strength is above the threshold required by the research purpose.

This contextual account would effectively deal with the problem raised to categorical accounts. But of course, it would fail to be an *unconditional* account of validity. Whether an instrument is valid or not would depend on the research purpose. Just like pragmatic encroachment theorists claim that to assess 'S knows P' we need to know how high a credence the specific context requires, this contextual account of validity claims that to assess "X is validly measured by this instrument" we need to know how much optimality the specific research purpose requires.

6.3. Graded Validity

However, there is a last-resource strategy available for those keen in defending an unconditional notion of validity: use the structure of ($b^{\text{Contextual}}$), but change the labels. They could say that we were talking inaccurately when we talked about outright validity. Validity is first and foremost a *graded* notion,

[12]As an analogy, consider pragmatic-encroachment approaches to knowledge. To define 'justified belief,' these accounts use two elements: a graded dimension of rational confidence in the proposition (the credence) and a context-dependent threshold (fixed by the stakes of believing the proposition, which depend on the purposes of the epistemic agent). The rational credence about *P* is *not* context-dependent. But an agent is justified in believing *P* iff her credence is above the context-dependent threshold.

not a categorical one—there is no such thing as an outright valid instrument. Rather, validity comes in degrees, and it is just what we have been calling here optimality (overall goodness). Degrees of validity just are degrees of optimality. And for this reason, proponents of this strategy can insist that validity—in its graded sense—is invariant to the research context. Whether instruments have higher or lower degrees of validity does not depend on research purposes. Only whether they have enough of it for grounding the inferences we are interested in—what I'm calling the *valid for* judgments—depends on the research purpose.[13]

In what follows, I argue that even this strategy fails. There is a further way in which context is required for judgments of validity, even of degrees of validity thus understood. Let's first illustrate the graded account with a possible case. Think of two instruments, A and B, that are causally affected by an existent attribute *X*. Instrument A has a specific degree of accuracy—say, it is accurate in measuring changes in *X* to the first decimal. Instrument B, in contrast, is accurate in measuring changes in *X* to the fifth decimal (i.e., B is more accurate, thus pro tanto more valid than A). Suppose, further, that in all other respects A and B are alike (e.g., the range of *X* they are able to measure is the same, their reliability, etc.). Thus, B has a higher level of validity than A. Imagine further that the judgments we are interested in require different levels of accuracy, but that with respects to other dimensions of optimality they all require the same very minimal levels. Those levels are satisfied by both instruments. It is clear here, then, that we can know whether these instruments are *valid for* the judgments we are interested in as long as we know the level of accuracy (with respect to the decimal position) that the different judgments require. Moreover, we know B is *valid for* all judgments A is *valid for*, and that A is *valid for* only some of the judgments that B is *valid for*.[14] Of course, under the graded account, there is no sense in claiming that B is outright valid or that A isn't. We can only say that B has a higher level of validity than A.

So far throughout the paper I have assumed that we can talk of higher and lower degrees of optimality unambiguously. This is surely required for a

[13]Stone (2019) also describes her account of "construct validity" as coming in degrees. In her account, however, the graded dimension is not validity, but the *evidence* we have for validity. That is, "more valid" means having more evidence for an instrument's validity. This is a consequence of the fact, mentioned above, that her account collapses validity with validation.

[14]Compare: a credence about *P* of 0.8 justifies believing in *P* in all contexts that a credence of 0.6 justifies, and a credence about *P* of 0.6 justifies believing in *p* only in some of the contexts where a credence of 0.8 justifies.

graded account of validity to make sense. But is this really plausible across the board? Is there a unidimensional quantitative scale—like in the case of credences—or at least an ordered and complete ranking, where we can situate *both* the level of validity that an instrument has for measuring *X and* the level of validity that is required by particular judgments? I argue here that this is not plausible across the board, so that the graded account of unconditional validity fails.

One situation in which this proposal would fail is as follows (a concrete example is given below). When (i) we have an instrument (instrument-A) that is *valid for* a specific judgment (judgment-1), but not for another judgment (judgment-2), and (ii) we have another instrument (instrument-B) that is *valid for* judgment-2 but not for judgment-1. This situation cannot occur within the graded account: this situation implies we cannot order instruments and judgments on the same level-of-validity scale. Hence, if this situation is possible, the graded account of validity fails.

A simple case of such situation is as follows. Take two instruments, A and B, which measure a dichotomic variable (e.g., 1 = depression, 0 = not-depression) in a given population. Instrument A fully satisfies the *all* part of (b^{Max}) but only partially satisfies the *only* part of (b^{Max}), such that, in the language used in medicine, A is 100% "sensitive" but not 100% "specific." Instrument B, in contrast, fully satisfies the *only* part of (b^{Max}), such that B is 100% "specific" but not 100% "sensitive." This means that, on some occasions of the given population, depression's presence is not detected by B and depression's absence is not detected by A.

Which instrument has a higher level of validity? The graded account must be able to answer this question without referring to the research purpose. However, there seem to be no good answer to the question of whether A or B has more validity independent of the research purpose. Given the differing strengths of A and B, A will be *valid for* some purposes for which B is not valid (e.g., those that need 100% sensitivity). One such purpose is ruling out the presence of depression in an individual. For this purpose, we know A is good enough, while B is not. So, if the graded account is correct, A must have a higher level of validity than B. However, B is *valid for* some purposes that need 100% specificity for which A is not valid—for example, for ruling out the absence of depression in an individual. If we now say that B has higher level of validity than A, we contradict ourselves.

It is, indeed, common talk in medicine that it is difficult to develop tests that are both highly sensitive and very specific. And the explanation for this seems straightforward. A measure of depression (such as HAMD) might include several questions, the responses of which are added up. In order to have a categorical outcome, a cut-score is needed. By using a higher or lower

cut-score, one increases specificity or sensitivity, but typically not both. (Think that using one extreme cut-score one can trivially assure 100% sensitivity, using the opposite extreme cut-score one can trivially assure 100% specificity.) Thus, the trade-off we are mentioning here is a rather common one. And the purposes of ruling out presence and absence of attributes are, plainly, not far-fetched. Therefore, the example provided here is no philosophical extravagancy. This example challenges the assumption of the graded view of validity—we are not able to order instruments A and B on the same level-of-validity scale without entering into contradiction. Fixing a research purpose would, of course, help us correctly judge which instrument has a higher level of validity *for that purpose*. But that amounts to rejecting the graded account of *unconditional validity*. Although I have no proof to offer, I cannot envision how a version of the graded account may be able to avoid this challenge. Any graded account will need to face the question of whether instrument A or B has a higher degree of validity: how can they provide an answer that does not depend on the research purpose?

This example is an instance that satisfies the following general set of conditions: (i) what we have been calling "overall optimality" (i.e., graded validity) is determined by various more specific optimality-dimensions (or validity-aspects), (ii) there are trade-offs between those specific dimensions, and (iii) the judgments relevant for different research purposes are more sensitive to different dimensions. These conditions jointly make it the case that some instruments are better for some purposes (because they perform better regarding the dimensions more relevant for those purposes), and other instruments perform better for other purposes (ditto).[15]

To sum up, the real possibility of trade-offs between goodness dimensions (such as sensitivity and specificity), and of different weights given to the different dimensions by different research purposes, entails that we are not always able to rank different instruments according to their degree of validity irrespective of the research purposes. But if we are not able to rank the instruments according to even their *degree* of validity irrespective of the research purposes, then, the very idea of unconditional validity is in question. The discussion of categorical accounts showed that we cannot always say whether an instrument is outright valid without stipulating a research purpose. And now we see that sometimes we cannot even say how valid an instrument is without stipulating a research purpose. The research purpose determines not

[15]These conditions are analogous to those that lead to a pluralism about models (e.g., Weisberg 2007, 656). There are various respects in terms of which models may be assessed, different models perform better in different respects, and different modelling purposes/aims value these diverse respects differently.

only how much validity is required, but also what *aspects* of validity (or optimality-dimensions) are more relevant.[16]

6.4. A Possible Objection?

It might be objected that the above discussion has overlooked a privileged source for finding a causal account of graded validity: the recent literature on measures of causal strength (Fitelson and Hitchcock 2011; Sprenger 2018). If degree of validity is determined by the strength of the causal relationship between the attribute and the outcomes of the measuring instrument, shouldn't we look for the preferred measure of causal strength in the literature to see whether a graded account of validity based on such a measure succeeds? Given that the problem raised for graded accounts of validity has to do with considering degrees of validity as not being unidimensional, perhaps basing the graded account with the single best measure of causal strength might defuse the problem. It might provide an objective way of balancing those dimensions, or, even better, a way of "reducing" those (apparently diverse) dimensions to only one.

Sprenger (2018) has persuasively argued that the preferred measure of causal strength is the one first suggested by Ellery Eells.[17] This measure defines the strength of the causal relation between cause C and effect E as follows:

$$\mathrm{CS}_{\mathrm{Eells}} = p(\mathrm{E} \mid \mathrm{C}) - p(\mathrm{E} \mid \sim \mathrm{C})$$

It seems natural to specify the degree of validity of different instruments with $\mathrm{CS}_{\mathrm{Eells}}$. Measurement outcomes that are not causally affected by the intended attributes would have a value of $\mathrm{CS}_{\mathrm{Eells}} = 0$; while those which are affected, and *only* affected, by the intended attributes would have a value of $\mathrm{CS}_{\mathrm{Eells}} = 1$. The larger the value of $\mathrm{CS}_{\mathrm{Eells}}$, the more valid the instrument. However, by using this measure to specify graded validity we run into the same problem raised above. To see this, we just need to define sensitivity and specificity in terms of E and C, assuming C and E to be dichotomous variables (as they are in our example above): $\mathrm{Sensitivity} = p(\mathrm{E}|\mathrm{C})$, and $\mathrm{Specificity} = p(\sim \mathrm{E}| \sim \mathrm{C})$.[18]
Because $p(\mathrm{E} \mid \sim \mathrm{C}) = 1 - p(\sim \mathrm{E} \mid \sim \mathrm{C})$, we can write $\mathrm{CS}_{\mathrm{Eells}}$ as follows:

[16]This last lesson implies that proponents of ($\mathrm{b}^{\mathrm{Contextual}}$) need to add the qualification that both the threshold *and* the degree to which *X*-variations cause *M*-variations are research-purpose-relative.

[17]Fitelson and Hitchcock (2011) also provide reasons to prefer Eells' measure.

[18]Note that here sensitivity and specificity are understood causally (versus in terms of correlations). In our case the "intervention" is the presence of the attribute (say, depression).

$$CS_{Eells} = \text{Sensitivity} + \text{Specificity} - 1$$

Plainly, Eells' measure of causal strength is *not* a measure of graded validity that would defuse the problem. If we have three instruments: one that is 100% sensitive but 0% specific, another that is 100% specific but 0% sensitive, and another that is 50% sensitive and 50% specific, they all have the *same* amount of Eells-causal strength. If we base the graded account of validity on Eells-causal strength, the three instruments are on a par—they are unconditionally valid to the same degree. However, as argued, depending on the research purpose some instruments will be *valid for* the relevant judgments while others will not. Decomposing Eells' measure of causal strength, as we did, shows that casual strength is not a concept more primitive than sensitivity and specificity. We cannot "reduce" the latter to the former. Rather, causal strength seems to be composed of (thus, reducible to) sensitivity and specificity. Also, Eells' measure gives equal weight to both sensitivity and specificity. Why would this be the correct weight across all research purposes? Given that these dimensions may not be equally relevant across all research purposes, there is no reason to believe that this measure provides us with an objective way of balancing these dimensions. Thus, considering the preferred measure of causal strength does not challenge our conclusion—it provides further support to our conclusion.

7. Conclusion

What it means for an instrument to validly measure a construct remains a lively debate. In this paper, I have not discussed several aspects of this debate—for example, whether validity should be defined in causal terms (vs. correlations); what kind of realism is, if any, required to make sense of measurement practices; how different accounts of validity lead to different preferred approaches to validation. My focus has been only in assessing the plausibility of causal accounts of *unconditional* validity. That is, accounts that are independent of context-specific aspects. I assessed several plausible candidates, finding all of them wanting. Any satisfactory account of validity, I argued, must restrict its judgments to context-specific situations. Thus, defining validity in causal terms does not avoid the need to restrict validity judgments to specific research purposes.

In light of the above, I suggest we should not look for general accounts of validity (or optimality). Instead, we should articulate the specific standards that are relevant for the attainment of our various specific goals. Rather than asking when are, say, depression surveys valid simpliciter, we should ask

under what specific conditions the measurements of a (less-than perfect) depression survey allows us to infer that depression is more prevalent in some groups than others, or allow us to judge as ineffective interventions that target symptoms while judging as effective those that actually target depression. The recent revival of the philosophy of measurement, I suggest, has yet to engage in these more context-specific projects. Nothing of what I have said suggests that a causal framework will not be of help for this task.

Acknowledgments

Thanks to Anna Alexandrova, Lukas Beck, Derek Briggs, Jacob Stegenga, and participants at the Philosophy of Social Science Roundtable (Atlanta, 2020) for helpful comments.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper is based on research that was funded by ANID (Chile), Gates Cambridge Trust, and St. Edmund's College (University of Cambridge).

References

Alexandrova, Anna. 2017. *A Philosophy for the Science of Well-being*. Oxford: Oxford University Press.

Alexandrova, Anna, and Daniel Haybron. 2016. "Is Construct Validity Valid?" *Philosophy of Science* 83: 1098-109.

Angner, Erik. 2011a. "Current Trends in Welfare Measurement." In *The Elgar Companion to Recent Economic Methodology*, edited by John B. Davis, and D. Wade Hands, 121-54. Northampton: Edward Elgar.

Angner, Erik. 2011b. "Are Subjective Measures of Well-being "Direct"?" *Australasian Journal of Philosophy* 89 (1): 115-30.

Borsboom, Denny. 2005. *Measuring the Mind: Conceptual Issues in Contemporary Psychometrics*. Cambridge: Cambridge University Press.

Borsboom, Denny, and Gideon Mellenbergh. 2004. "Why Psychometrics is Not Pathological: A Comment on Michell." *Theory and Psychology* 14 (1): 105-20.

Borsboom, Denny, Gideon Mellenbergh, and Jaap Van Heerden. 2004. "The Concept of Validity." *Psychological Review* 111: 1061-71.

Cartwright, Nancy and Norman Bradburn. 2011. "A Theory of Measurement." In *The Importance of Common Metrics for Advancing Social Science Theory and*

Research: Proceedings of the National Research Council Committee on Common Metrics, edited by National Research Council, 53-70. Washington, DC: National Academies.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. Oxford: Oxford University Press.

Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (4): 281-302.

Davidson, Donald. 1969. "The Individuation of Events." In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, 216-34. Reidel.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth. 2016. "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution." *American Economic Journal: Applied Economics* 8 (4): 36-65.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89 (2-3): 381-94.

Fitelson, Branden, and Christopher Hitchcock. 2011. "Probabilistic Measures of Causal Strength." In *Causality in the Sciences*, edited by Phyllis McKay Illari, Federica Russo, and Jon Williamson, 600-27. Oxford: Oxford University Press.

Graff Zivin, Joshua S., Solomon Hsiang, and Matthew Neidell. 2018. "Temperature and Human Capital in the Short and Long Run." *Journal of the Association of Environmental and Resource Economists* 5: 77-105.

Gulliksen, Harold. 1950. *Theory of Mental Tests*. New York: Wiley and Sons.

Hacking, Ian. 1983. *Representing and Intervening*. New York: Cambridge University Press.

Haslanger, Sally. 2000. "Gender and Race: (What) Are They? (What) Do We Want Them to Be?" *Noûs* 34 (1): 31-55.

Hitchcock, Christopher. 1992. "Causal Explanation and Scientific Realism." *Erkenntnis* 37: 151-78.

Hood, S. Brian. 2009. "Validity in Psychological Testing and Scientific Realism." *Theory & Psychology* 19 (4): 451-73.

Hood, S. Brian. 2013. "Psychological Measurement and Methodological Realism." *Erkenntnis* 78 (4): 739-61.

Kane, Michael T. 2006. "Validation." In *Educational measurement* (4th edition), edited by Robert L. Brennan, 17-64. Westport, CT: Praeger.

Kelley, Truman L. 1927. *Interpretation of Educational Measurements*. New York: Macmillan.

Larroulet Philippi, Cristian. "Against Prohibition (Or, When Using Ordinal Scales to Compare Groups is OK)." Unpublished manuscript.

Markus, Keith A., and Denny Borsboom. 2013. *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. Routledge.

Messick, Samuel. 1989. "Validity." In *Educational Measurement* (3rd edition), edited by Robert L. Linn, 13-103. Washington, DC: American Council on Education and National Council on Measurement in Education.

Michel, Matthias. 2019. "The Mismeasure of Consciousness: A Problem of Coordination for the Perceptual Awareness Scale." *Philosophy of Science* 86 (5): 1239-49.

Northcott, Robert. 2013. "Verisimilitude: A Causal Approach." *Synthese* 190 (9): 1471-88.

Paul, Laurie A., and Ned Hall. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Sprenger, Jan. 2018. "Foundations of a Probabilistic Theory of Causal Strength." *The Philosophical Review* 127 (3): 371-98.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Stevens, Stanley Smith. 1946. "On the Theory of Scales of Measurement." *Science*, 103 (2684): 667-80.

Stone, Caroline. 2019. "A Defense and Definition of Construct Validity in Psychology." *Philosophy of Science* 86 (5): 1250-61.

Tal, Eran. 2017. "Measurement in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/measurement-science/

Vessonen, Elina. 2017. "Psychometrics Versus Representational Theory of Measurement." *Philosophy of the Social Sciences* 47 (4-5): 330-50.

Vessonen, Elina. 2019. "Operationalism and Realism in Psychometrics." *Philosophy Compass* 14 (10): e12624.

Weisberg, Michael. 2007. "Three Kinds of Idealization." *The Journal of Philosophy* 104 (12): 639-59.

Author Biography

Cristian Larroulet Philippi is a PhD Candidate at the Department of History and Philosophy of Science, University of Cambridge. His dissertation tackles different challenges about measurement. Before studying philosophy, he studied and did research in economics. He has previously published on values in science.