Conceptual engineering and operationalism in psychology

Elina Vessonen¹

Received: 26 October 2020 / Accepted: 11 June 2021 / Published online: 14 July 2021
© The Author(s) 2021

Abstract

This paper applies conceptual engineering to deal with four objections that have been levelled against operationalism in psychology. These objections are: (i) operationalism leads to harmful proliferation of concepts, (ii) operationalism goes hand-in-hand with untenable antirealism, (iii) operationalism leads to arbitrariness in scientific concept formation, and (iv) operationalism is incompatible with the usual conception of scientific measurement. Relying on a formulation of three principles of conceptual engineering, I will argue that there is a useful form of operationalism that does not fall prey to these four objections.

Keywords Operationalism · Conceptual engineering · Concept formation · Rudolf Carnap · Psychology

1 Introduction

Conceptual engineering and operationalism are linked by content and by history. In terms of content, both are concerned with the appropriate formulation of (scientific) concepts. While the details and the domain of applicability of conceptual engineering are debated, it is agreed that conceptual engineering involves the creation and revision of concepts so as to make them appropriate for a given task (Cappelen, 2018). For instance, an appropriately engineered concept of truth might be such that it performs relevant explanatory work while avoiding paradoxes that traditional concepts of truth tend to cause (Scharp, 2013). Operationalism, on the other hand, is concerned with the formulation of concepts that are appropriate for a specific task, namely, measurement or testing (Chang, 2009). For instance, the operationalization of the concept of well-being requires characterizing well-being so that it is defined

This article belongs to the topical collection “Concept Formation in the Natural and Social Sciences: Empirical and Normative Aspects”, edited by Georg Brun, Catherine Herfeld, and Kevin Reuter.

✉ Elina Vessonen
elina.vessonen@gmail.com

¹ Visiting Researcher At Finnish Institute for Health and Welfare, PL 30, 00271 Helsinki, Finland

in terms of a measurement operation or otherwise measurable terms—the choice between these options, and the ontological status of the resulting concept depends on how strict a form of operationalism one subscribes to.

As for the historical link, both operationalism and conceptual engineering have connections to logical empiricism and Rudolf Carnap. The connection to logical empiricism is most obvious in the case of conceptual engineering. Most authors trace the origins of conceptual engineering to the work of Rudolf Carnap—indeed the term “conceptual engineering” reportedly first appears in Richard Creath’s introduction to a book containing correspondence between Carnap and Willard van Orman Quine (Creath, 1990). In his 1950 book, Carnap sketched the method of explication, which involves the re-characterization of scientific concepts by balancing four criteria a useful concept ought to fulfill (in Carnap’s view): exactness, similarity to prior usage, fruitfulness and simplicity (Carnap, 1950b). The method of explication is regularly viewed as an example of, or a precursor to conceptual engineering (e.g. Scharp, 2013).

The connection between operationalism and Carnap is more indirect. The first and the most famous explicit characterization of operationalism is due to Nobel Prize winning physicist Percy W. Bridgman, whose ideas on the operational definition of scientific concepts drew the interest of experimental psychologists and logical empiricists in the 1930s and 1940s (Chang, 2009; Green, 1992).¹ Bridgman’s focus on concepts defined in testable and measurable terms resonated with logical empiricists, some of whom weighed in on debates about operationalism alongside Bridgman and experimental psychologists (Hempel, 1954). While Carnap did not contribute explicitly to these debates, he seems nonetheless to have been on the radar when operationalism was formulated and adopted in psychology (Stevens, 1935).

Notwithstanding these connections and similarities (which certainly do not exhaust the relevant historical links), conceptual engineering and operationalism have had very different trajectories in academic research, and specifically within philosophy. Operationalism received quite a bit of attention in its heydays in the 1930s and the 1940s, and arguably had a significant impact on the way researchers started to approach the measurement of psychological attributes (Michell, 1990). But the main participants to debates on operationalism—Bridgman, a handful of logical empiricists and experimental psychologists—found fatal flaws in each other’s formulations of operationalism and were unable to agree on how the approach should be applied. By mid-1950s, Bridgman felt that he had “created a Frankenstein” (Bridgman in (Frank, 1956), quoted in (Chang, 2009)), and most philosophers abandoned operationalism, perhaps partly because it was viewed as an extension or a close relative of logical empiricism (Green, 1992).

Although several authors have claimed that contemporary psychological measurement continues to proceed along operationalist lines in *practice* (Lovett & Hood, 2011; Maul & McGrane, 2017), operationalism *qua philosophy* has fallen in disrepute. Accordingly, we find contemporary authors writing that operationalism has

¹ It can be argued that Peirce (1878) and Watson (1913) characterized something similar without using the name operationalism.

been “almost uniformly rejected as philosophically unworkable” (Maul, 2017, p. 60) and that operationalism is an “erroneous philosophy of science” (Meehl, 1995, p. 267). Two notable exceptions to these condemnations are the works of Hasok Chang (2009, 2017) and Uljana Feest (2010, 2012), who argue that when appropriately interpreted, operationalism continues to have value for philosophy as well as for science.

Conceptual engineering, by contrast, has been a silent—or at least an unnamed—approach until very recently. Although the approach was reportedly named already in the 1990s (Creath, 1990), and something like it has probably been around way before Carnap explicated explication, it is only in the last decade or so that conceptual engineering has become a popular topic in metaphilosophy. One reason for this might be that philosophers have been more interested in analyzing concepts than revising them (or at least it has been common for philosophers to conceive of their activities in these terms). Accordingly, conceptual analysis has been thought of as *the* method of analytic philosophy (e.g. Scharp, 2013). In recent years, conceptual engineering, and the closely related ameliorative analysis (Haslanger, 2000), have been explicitly applied to various concepts such as truth and gender.² With increased attention, various challenges and criticisms have also been levelled against conceptual engineering. It is still fair to say that conceptual engineering is not yet a fully formed, single methodology.

So, we have two frameworks that are broadly speaking related to concept formation, one of which has fallen in disrepute, while the other one has only recently become the target of rigorous formulation, investigation and debate. In this paper I am going to discuss the relationship between these two frameworks, operationalism and (Carnapian) conceptual engineering. Focusing on operationalism as it manifests in psychology, I’ll argue that many of the major challenges levelled against operationalism can be reasonably countered if we adopt the perspective of conceptual engineering. In other words, central aspects of (a particular form of) conceptual engineering help redeem (a particular form of) operationalism. This has implications for contemporary psychology, because psychological measurement is frequently thought to have problematic operationalist commitments. Although conceptual engineering might not be necessary for redeeming operationalism, I argue that it does offer one compelling way to defend operationalism in psychology. To my knowledge, the implications of conceptual engineering on psychological operationalism have not been explored before.

I mentioned above that I will apply *a particular form of* conceptual engineering to counter certain criticisms against operationalism. What if one does not buy this formulation of conceptual engineering? While my discussion will not satisfy such a person that (some of) the problems of operationalism have now been resolved, the discussion at least helps us pinpoint metaphilosophical sources of disagreement about operationalism. In other words, we will be able to trace a disagreement about operationalism to a more general, methodological dispute about concept formation. This might help us avoid talking past each other in debates about operationalism.

² For a comparison of the two approaches, see Dutilh Novaes (2020).

The structure of the paper is the following. Section 2 describes operationalism. Section 3 enumerates four common objections to operationalism. Section 4 describes conceptual engineering. Section 5 contains an analysis of operationalism in terms of conceptual engineering. Section 6 concludes with remarks on how the present argument helps contemporary psychology.

2 Operationalism

2.1 What is it?

Operationalism stands for multiple things (Vessonen, 2021). For some it is a theory of meaning: it tells us what it means to say that “Maya is depressed”. For others it is a research heuristic: it tells us how to study whether or not Maya is depressed. For yet others operationalism is a metaphysical thesis: it tells us what depression is, for example, that depression is a bodily condition (cf. Flanagan, 1980). Because operationalism has many meanings, it is easy to misunderstand the type of operationalism a particular author is defending or critiquing. Indeed, there are papers arguing that S. S. Stevens’ operationalism was widely misunderstood (Feest, 2005) and that Bridgman did not subscribe to the views that are often attributed to him (Chang, 2017).

Due to their large number, I am not going to go over all possible formulations of operationalism here. Rather, I will focus on operationalism *qua* the belief that some scientific concepts can and should be defined in terms of a test operation.³ For example, intelligence can be defined in terms of an IQ-test, depression in terms of an interviewer-administered rating scale such as the Hamilton Depression Rating Scale, and well-being in terms of a life satisfaction questionnaire.

This rendition might strike as too lenient or local to count as genuine operationalism. Some scholars think that operationalism denotes a universally applied position, which says that all scientific (or psychological) concepts should always be defined in terms of operations. As mentioned, in the history of science, various methodologies and philosophies have been called operationalism (Chang, 2009; Feest, 2005). Who is to say which definition of operationalism is the correct one, or the one we ought to be interested in? I don’t think there is any way to identify *the correct* definition of operationalism, but I do have a suggestion for what constitutes a notion of operationalism that is worth discussing. Firstly, when the aim is to understand and improve present-day psychological methodology, the focus should be on forms of operationalism that actually occur in scientific practice. Second, given these aims, we should not waste time on forms of operationalism that are so trivial that no one objects to them. I choose to focus on what one might call local operationalism here—the belief that scientific concepts can and sometimes should be defined in terms of a test operation—for just these reasons. As I will argue below, local operationalism is actually practiced in psychology. Therefore, an analysis of local operationalism matters for

³ The exact definition of an operation is debated. In this paper we can focus on relatively obvious cases such as paper-and-pencil tests. See Bridgman (1959) for a classification.

real life science. Second, local operationalism does face objections, as I will show in Sect. 3. Hence there is actually something to defend here. Thus, in this paper operationalism denotes local operationalism.

There is another form of psychological operationalism with which I should juxtapose the above described local operationalism. Feest (2005) argues that some of the best-known operationalist psychologists, S. S. Stevens and Edward Tolman, subscribed to a form of operationalism that does not fall prey to the typical objections to operationalism. The notion of operationalism that Feest analyses is different from the one I focus on here. Hence Feest's case, while important, does not salvage local operationalism as described above. Local operationalism needs a defense of its own.

According to Feest (2005, p. 133), when Stevens and Tolman devised operational concepts, they were “partially and temporarily specifying their usage of certain concepts by saying which kinds of empirical indicators they took to be indicative of the referents of the concepts.” I take this to mean that operationalists' concepts have test-independent referents—e.g. the concept of well-being refers to something that is independent of us testing it—and the purpose of operational definitions is to specify the experimental set up in which that referent manifests itself. Feest (2005, p. 136) elaborates with reference to Stevens' contention that “to experience is, for the purpose of science, to react discriminately”:

“Did he mean by this that the expression experience has the same meaning as discriminative behavior? Did he mean that the presence of discriminative behavior is always a necessary and sufficient condition for the correct application of the term experience? Based on his research, I think that this is clearly not what he has in mind. [...] The question, for him, is how to “get at” particular kinds of phenomenal experience in an experimental context. His assertion is that this can only be done via the behavior of the organism—i.e., that in an experiment, discriminative behavior is a necessary condition for attributing conscious experience to an organism.”

Feest's analysis of Stevens' operationalism, and its resistance to common objections, is compelling. But this notion of operationalism, which Feest calls methodological operationalism, is different from the one I am dealing with here. If Stevens *had* been a local operationalist in my sense of the word, he would have subscribed to the view that experience *does* mean discriminative behavior. Or in terms of a specific “experience”: a local operationalist would say that e.g. happiness *means* responding in a certain way to a happiness questionnaire. Here the operational definition does not refer to a test that captures the referent of the concept of happiness, but rather the referent of the concept of happiness is behavior in response to the test. Later we will see that the local operationalist can accept the co-existence of various characterizations of happiness, but for the purposes of research, the local operationalist does reduce happiness to a test.

Why would anyone defend local operationalism, if the philosophically less demanding methodological operationalism is available and can be shown to be useful? It seems to me that methodological operationalists face epistemic issues that other forms of operationalism avoid. In particular, if operational definitions encode the experimental set-ups in which referents of the target concept manifest, it can

always be asked whether the set-up really does capture the intended referent. For example, if one were to state that a specific IQ test “gets at” the referent of the concept of intelligence, it is legitimate to demand evidence that the test really does get at that referent. Whereas if intelligence is defined in terms of an IQ test, there is no epistemic gap between the referent of the concept of intelligence and the IQ test—the referent of the concept of intelligence is nothing more than responses to an IQ test. Local operationalism allows safe passage from test results to claims about the target concept by defining away the gap that separates them in methodological operationalism.

2.2 Local operationalism in the wild

Practicing research psychologists rarely take any explicit stance on operationalism. Rather, they routinely engage in the act of operationalization, which is described in many textbooks of psychological measurement. In their textbook, Evans and Rooney (2011) describe operationalization as an activity that turns a conceptual hypothesis into a research hypothesis. A conceptual hypothesis is an intangible claim such as “Outgoing people have higher well-being than people who keep to themselves” and a research hypothesis is a claim about test outcomes such as “People who score high on standard test E of extraversion give higher ratings of well-being on test W than do people who score low on E.” Here “ratings on test T” is an operational conception of well-being, because it defines well-being in terms of a test, which is a type of scientific operation.

The fact that a researcher operationalizes happiness in terms of a rating on test H does not imply that the researcher believes there is nothing more to happiness than the way people respond to test W. One common, epistemically innocent way to make sense of the act of operationalization is this: the operational concept of well-being gets linked back to the “true” concept of well-being when researchers *validate* test W. For instance, if the true concept of well-being is “frequent positive affect and infrequent negative affect”, then the researcher needs to ensure that test W tracks this concept by validating the test. Understood in this way, operationalization amounts to making the intended, non-operational target concept measurable, rather than reducing the concept of interest to a more superficial, test-specific concept.

Now, it is true that psychologists routinely validate their measures: they check correlations between the proposed test and tests of related concepts, study relations between different items (i.e. questions) within the test, and perform factor analysis or an item response theoretic analysis to model the structure of responses to the measure (for an overview of these methods, see Markus & Borsboom, 2013; Nunnally & Bernstein, 1994).⁴ But does this guarantee that the measure tracks the intended, non-operationally-defined target concept? Many psychometricians, methodologists and philosophers of science have argued that the typical methods of validation in psychology are simply insufficient to establish that a psychological test tracks the concept of interest

⁴ I shall use the term validation in this broad sense throughout the paper.

(Alexandrova & Haybron, 2016; Borsboom, 2006; McClimans et al., 2017). For example, Andrew Maul (2017) showed in a series of studies that fake psychological tests, that is, tests that e.g. ask questions about non-existent psychological properties, would be classified as valid under common criteria for assessing the properties of a test. Why might tests fall short of tracking a test-independent attribute? Maul et al. (2013) argue that when the methods of validation that are routinely used in contemporary psychology were originally formulated, their creators were heavily influenced by logical empiricism and behaviorism. Some validation practices might not allow inferences to the non-operational concept of interest, because they were not even intended to warrant these kinds of inferences.

Whatever the reason for their existence, what do the shortcomings of validation methods imply regarding the role of operationalism in psychology? If the validation methods are insufficient for ensuring that the measure tracks a non-operational concept, then claims made on the basis of those measures are superficial claims about test-behavior, not about the concept of interest. In other words, claims about the relation between well-being and outgoingness would simply be claims about two kinds of test results. It may be that this slip from claims about non-operational concepts to claims about operational concepts is sometimes accidental: a researcher is unaware of the fact that the validation does not establish a sufficient link between the test and the non-operational concept of interest. On the other hand, it may be that in the absence of (knowledge about) superior methods, research psychologists have simply accepted that the best one can do is to characterize the target concept in terms of a test.

Above I have presented two interpretations of how operationalism might slip into psychology: by accident or as an inevitable consequence of lack of (knowledge about) better methods. But there are other, more radical interpretations as well. For instance, Joel Michell (1997, 2008) has criticized psychological measurement extensively, arguing that ever since S. S. Stevens popularized operationalism, psychologists have conceptualized measurement in operationalist terms. In other words, operationalism is not an accidental slip, but rather the rewriting of non-operational concepts in operational terms is the accepted framework for psychological measurement. For the purposes of this paper, we need not settle once and for all the manner in which operationalism slips or arrives to psychology. It suffices to note that psychologists at least sometimes make claims where a concept that has a broad, non-operational meaning gets reduced to an operationally defined concept.

3 Objections to operationalism

In this section I will introduce some common objections levelled against operationalism, focusing on criticisms of the type of operationalism that is (allegedly) practiced in psychology. The four objections I consider are:

- (i) operationalism leads to harmful proliferation of concepts,
- (ii) operationalism goes hand-in-hand with untenable anti-realism,
- (iii) operationalism leads to arbitrariness in scientific concept formation, and

- (iv) operationalism is incompatible with the usual conception of scientific measurement.

3.1 Proliferation

In debates about operationalism, one often hears the charge that operationalism leads to harmful proliferation of concepts (e.g. Hempel, 1966; Hull, 1968; Leahey, 1980; Maul et al., 2013). If different operations each define a new concept, scientists will drown in concepts—or so the objection goes. The prospect of drowning in concepts sounds odd and undesirable on its own, but it also has more tangible, harmful consequences. Firstly, it is inefficient to produce more and more concepts in situations where one or two could do the job. Second, an ever-increasing stock of operational concepts is a threat to comparability. If researchers have two tests of well-being, and one of them is strongly positively correlated with income while the other is not, are they to conclude that well-being both is and is not associated with income? They might have to accept that conclusion, if they believe that both tests define a distinct concept of well-being.

3.2 Antirealism

It is often thought that operationalism is an antirealist view (Lovett & Hood, 2011; Maul & McGrane, 2017).⁵ It is, in other words, thought that operationalists deny the test-independent existence of the attributes their measures apparently pertain to and/or they deny the possibility of epistemic access to a test-independent attribute. On this view, for example, operationalists define depression in terms of a test because they believe either (i) that there is nothing more to depression than the way people behave in depression tests (i.e. there is no independent cause to these symptoms), or (ii) that there is no way to access the test-independent determinant of that behaviour. By test-independent attribute I mean an attribute that can be fully characterized without mentioning the operation used to measure it. For example, *cognitive assessment of satisfaction with one's life* (whether or not it is observed) is a test-independent attribute, and the concept that denotes that attribute is a non-operational concept. On the other hand, *self-reports in response to the Satisfaction with Life Scale* is a test-dependent attribute (or property, or quality, whatever term one likes), and it is denoted by an operational concept.

There are, of course, philosophical defences of antirealism. But many people believe that science, including psychology, ought to study entities and phenomena that exist (in some sense) independent of scientists' tests and measures. To such people, information about how people answer to questions about their moods seems like a pale and useless shadow compared to substantive knowledge about what drives those test responses, e.g. depression *qua* independent, potentially causally

⁵ Antirealism (and realism) comes in many forms but I will focus on the styles most relevant to operationalism, as described in (Lovett & Hood, 2011). See Trout (1998) on types of realism.

efficacious aspect or property of the human mind. Of course, it might be that some terms psychologists use do not have any test-independent referents—perhaps there is no unified, causally efficacious attribute or process that underlies the symptoms that depression tests inquire about. But that is a matter of investigation. To “go operationalist” (and thereby antirealist) prior to such an investigation is unnecessary defeatism, the critic would say.

3.3 Arbitrary concepts

The third critique flows from the second. According to this objection, if some of the targets of psychological research—happiness, depression, personality and so on—have no test-independent reality, then nothing constraints the formation of those concepts, or the construction of the measures meant to capture those concepts. Another way to state this problem is to say that if there is no truth about what, say, well-being is, then nothing stops researchers from defining well-being in whatever (operational) way pleases them. Andrew Maul and Joshua McGrane make this argument in their 2017 paper. In response to Hand (2016), who argues that the operationalist underpinnings of (some parts of) psychological measurement do *not* imply that anything goes, Maul and McGrane (2017, p. 2) assert that “when measurement is freed from any reality constraints, the inevitable outcome is (as repeatedly demonstrated throughout the social and psychological sciences and Hand’s volume) that anything *does* go.”

An anything goes -approach to concept formation would certainly be a disaster. Most importantly, if researchers can define well-being and intelligence in any way they want, then they can also force almost any empirical claim to come out true. Well-being measure W fails to correlate with income in a convenient manner? No problem! Just revise the measure, and the corresponding operational concept of well-being, until the desired correlation emerges. This is ideology, not science. Furthermore, unconstrained concept formation certainly exacerbates problem number i), concept proliferation.

3.4 Non-measurement

Several authors have also noted that operationalism seems to be unable to accommodate common notions of what measurement is or requires. For instance, Donald Gillies (1972) argued that strict operationalists cannot make sense of claims such as “Type A thermometers measure temperature more accurately than Type B thermometers”. If every measure defines its target concept, there is no way scientists can compare two measures on their ability to track common target values. Put in another way, operationalism seems unable to accommodate the idea that some measurement results “contain” more error than others. The objection, then,

is that an approach to measurement that cannot make sense of error and accuracy is simply not science (Michell, 1990).

This objection is interesting, because it reveals a tension or an inconsistency in psychological measurement: on the one hand, it is characterized by operationalist tendencies; on the other hand, psychologists take steps to analyze and reduce sources of error in measurement. If, as per objection ii), operationalism is an anti-realist view, then psychologists' efforts to get rid of error are puzzling. If there is nothing scientists' measures are trying to represent or track correctly, what is there to be in error about?⁶

4 Three aspects of Carnapian conceptual engineering

Many challenges have been levelled against conceptual engineering in recent years. Due to the open questions and as yet unanswered criticisms, there are various ways to conceptualize conceptual engineering, and not all of them are compatible with each other. In this section, I am going to outline three aspects of conceptual engineering that I draw from the work of Rudolf Carnap (for a thorough discussion of Carnapian explication see Brun, 2016).

Here are the three Carnapian insights I use to try to rescue operationalism from some of its critics:

- (a) concept formation requires the balancing of (epistemic) values;
- (b) concept formation is constrained both from the side of human interest and from the side of reality; and
- (c) concept formation is an iterative process involving conceptual pluralism.

As mentioned in the introduction, I think the analysis that follows is valuable even if the reader finds one or more of these three principles controversial. In that case, while my discussion will not satisfy such a person that the problems of operationalism have now been resolved, the discussion allows us to pinpoint the metaphilosophical source of disagreement about operationalism. For instance, in Sect. 5 I will argue against objection (iii), which concerns the alleged anything goes -nature of operationalist concept formation, by appealing to principle (a), which states that concept formation involves the balancing of epistemic values. This allows us to pinpoint at least one reason for why researchers disagree on whether operationalism is an anything goes -methodology. But before re-analyzing operationalism, let me introduce each conceptual engineering principle in turn.

⁶ Borsboom, Mellenbergh and van Heerden (2003) make a similar point when they argue that a psychometric activity known as latent variable modelling cannot be understood in operationalist terms (albeit they use the word constructivism).

4.1 Values in concept formation

When Carnap (1950b) laid out his method of conceptual engineering, he suggested that the re-characterization of inexact, pre-theoretic or currently used scientific concepts should be done in light of the following criteria: similarity to the pre-existing concept, exactness, fruitfulness and simplicity. The similarity criterion says, in effect, that there must be some continuity between the connotations and/or denotations of the pre-existing concept and those of the proposed explication—otherwise the exercise would amount to changing the subject (see Strawson, 1963 and Carnap’s response in the same volume). Exactness means connecting the new concept to a network of pre-existing concepts by means of a definition or another mode of characterization. Fruitfulness, arguably the most important of the criteria, requires that the concept is useful for the formulation of scientific generalizations. I call these criteria epistemic values, because they specify properties of concepts scientists tend to find valuable.

While these criteria might have sufficed for Carnap’s project of explicating probability, they may need to be complemented or replaced by other values in other contexts. For instance, in sciences that deal with thick or value-laden concepts, such as the science of well-being, scientists might need criteria such as impartiality or avoidance of value-imposition to form an appropriate concept (Alexandrova, 2016). Likewise, some authors have argued that while many scientific concepts are not useful for generalizations, they are still fruitful in other ways (Shepherd & Justus, 2015). Luckily, we need not settle which precise values are acceptable in the context of concept formation. It suffices to note that conceptual engineers balance different values when formulating new concepts.

4.2 Preferences meet reality

The Carnapian focus on values might make concept formation look like an activity that is all about what scientists want from their concepts, not about what reality,⁷ or empirical evidence that testifies for reality, says about those concepts. And indeed, Carnap often expressed his belief that people should tolerate various linguistic or conceptual systems, and grant scientists “*the freedom to use any form of expression which seems useful to them*” (Carnap, 1950a, p. 40). Is conceptual engineering precisely the kind of antirealist arbitrariness that operationalists seem to engage in?

The role of empirical evidence comes to view when we zoom in on the word “useful” in the above quotes, and interpret it in terms of Carnap’s criteria of explication. Remember that according to Carnap, a scientific concept should be fruitful in

⁷ By reality I shall mean the things, attributes and processes that push back when scientists use empirical means to probe the nature of the extension of a proposed concept. In other words, reality is what signals when scientists ask the internal question: Given that I have defined concept C in terms of X, what can I say about the denotation of C on the basis of the empirical evidence? Following Salmon (1994), I believe that these kinds of questions would have been acceptable and meaningful to Carnap, and that it is therefore not wrong to call my branch of conceptual engineering “Carnapian”. On external and internal questions, see (Carnap, 1934).

the sense that it is useful for the formulation of many empirical generalizations (or logical theorems) (1950b, p. 7). But what does it take for a concept to be useful for the formulation of such generalizations? Presumably the extension of that concept should have causal or other regular relationships with other entities or attributes. For instance, suppose that a particular definition of depression says that depression is a collection of five different mood-related symptoms. This concept allows us to formulate (non-trivial) generalizations, if reliable methods of investigation show robust relations between the specified depressive symptoms on the one hand and, say, proposed causes and treatments of those symptoms on the other. In other words, confirming that this concept of depression is fruitful requires empirical investigation of relations between the denotation of the proposed concept and other entities, attributes and processes.⁸

In the Carnapian vision, then, while scientists formulate and choose concepts according to their judgment of the fulfillment of certain epistemic values, those judgments are informed by, and revised in light of empirical evidence about the properties and behavior of the extension of the proposed concept. But one might still ask: which one has the upper hand? If scientists' preferences do not cohere with the results of empirical enquiry, shouldn't truth prevail over preferences?

Now, there are many ways in which the results of empirical enquiry and the conceptual preferences of scientists rub against each other. I cannot provide a manual for dealing with each situation here. But let me note that empirical evidence often leaves room for various conceptual choices. For instance: suppose one believes that a collection of symptoms counts as a disease only if those symptoms are caused by a unified biological mechanism. Suppose further that after decades of searching, researchers cannot find any such mechanism behind depressive symptoms, but rather, the symptoms seem to be caused by a messy network of social and biological processes. Does this mean that empirical evidence has shown that depression cannot be conceptualized as a disease—that it would be unscientific to believe that depression is a disease?

Well, no. As Quine taught decades ago, one can make legitimate adjustments in various nodes of the web of scientific beliefs. In this case, researchers can revise the notion that a unified biological mechanism is a necessary condition for something to count as a disease. Or they can limit the number of symptoms they count as essential to depression and see if they can find a mechanism that underlies variation in those symptoms. There is nothing about the above-described empirical results that prohibits either of these conceptual moves. To decide which move is the most appropriate, scientists need to consider other evidence and trade-offs between epistemic values: Would the narrower concept of depression be too dissimilar to the pre-theoretic notion of depression? Would the broadened concept of disease be too complicated?

⁸ A philosophical assumption behind this description is that characterizations of concepts usually do not exhaust all stable properties of the denotation of that concept. For instance, the concept "an overall cognitive judgment of how life is going for an individual" denotes a certain process or attribute, but it is still possible to be surprised by empirical evidence about that process—say, the neural phenomena underlying that judgment, the effect of age on that judgment, and so on.

Would disease cease to be a fruitful concept? As these considerations illustrate, empirical evidence and human judgment are inextricably intertwined in conceptual engineering.

4.3 Iteration and pluralism

The above outlined interplay of explications and empirical investigations fits neatly with Carnap's idea of concept formation as a stepwise, iterative process. In the 1950 book, he writes:

"[...] the historical development of the language is often as follows: a certain feature of events observed in nature is first described with the help of a classificatory concept; later a comparative concept is used instead of or in addition to the classificatory concept; and, still later, a quantitative concept is introduced." (1950b, p. 12).

Carnap uses the concept of temperature as an example. He suggests that temperature was initially a qualitative concept denoting sensation-based judgments of hot and cold but then gradually developed into the present-day quantitative concept via a comparative conceptualization. In this process, the new concept is built on the old one, in the sense that it retains some of the meanings or usages of the prior concept, but also improves it in terms of certain epistemic values, especially fruitfulness. While Carnap does not make reference to historical evidence, the philosophico-historical work of Hasok Chang (2004) has shown that the development of the scientific concept of temperature can indeed be understood as a kind of iterative process of self-correction and enrichment of earlier concepts.

Now, it is unrealistic to think that all scientists in a given field at a given time subscribe to one characterization of the central concept of that field. It is equally unrealistic to think that they revise that characterization synchronously in exactly the same way. Instead, a scientific field often contains a plurality of conceptions that are denoted by the same term and that share some connotations. For instance, well-being has been conceptualized both in terms of preference satisfaction and in terms of the ratio of positive to negative affect. While these concepts are different and would likely lead to different kinds of empirical generalizations, they are both explications of well-being, that is, of what is good for a person (Tiberius, 2006).

This kind of plurality of side-by-side existing, potentially competing conceptualization is not only a reality, but also likely to occur whenever Carnapian conceptual engineering is exercised. This is because scientists are likely to make trade-offs between epistemic values in different ways, leading them to propose and work with different kinds of explications. For instance, to avoid imposition of values, a researcher might need a rather intricate, multidimensional concept of well-being, while a scientist who needs to measure well-being on a quantitative scale probably prefers a unidimensional concept. Furthermore, different researchers might take different aspects of the shared pre-theoretic notion as their raw material when explicating a scientific concept. For example, one researcher may focus on the centrality

of low mood in lay conceptions of depression, while others regard complexity and multidimensionality as a key aspect of depression.

The upshot is that when practicing conceptual engineering, scientists are likely to have to deal with a plurality of side-by-side existing explications, each indexed to the same or similar term. But how does one manage a multitude of concepts? Carnap believed in a kind survival of the fittest when it comes to navigating and managing conceptual pluralism. He wrote:

“Let us grant to those who work in any special field of investigation the freedom to use any form of expression which seems useful to them; the work in the field will sooner or later lead to the elimination of those forms which have no useful function.” (1950a, p. 40).

While it is too optimistic to think that this process never misses a useful concept, or that it always converges on the most useful concept(s), iterative revisions and the assessment of concepts in terms of epistemic values puts at least some breaks on the expansion of the library of concepts.

With this final piece of the conceptual engineering framework in hand, we are ready to apply these insights to operationalism. In the next section, I will go over the four problems I identified earlier and argue that thinking in terms of conceptual engineering helps resolve criticisms levelled against operationalism. While it remains the case that not all forms of operationalism can be defended (at least not by the below arguments), a conceptual engineering approach reveals a form of operationalism that does not crumble under the most common criticisms.

The following discussion also reveals that contemporary psychology already contains measure validation practices that would allow researchers to implement the principles of conceptual engineering. I will use these practices as examples of how operationalist conceptual engineering would work in practice. But let me emphasize that these examples do not mean that contemporary psychologists already think in terms of conceptual engineering or the principles outlined above. As mentioned earlier, there is considerable disagreement regarding what different validation methods are intended to establish and what they manage to establish in reality. I will discuss the implications of conceptual engineering for contemporary psychology in Sect. 6.

5 Conceptual engineering redeems operationalism

5.1 Values, proliferation and comparability

The first charge against operationalism was that it leads to uncontrollable proliferation of concepts. Isn't it terribly inefficient for each measure to define its own concept? How can scientists ever compare such measures? In the conceptual engineering edition of operationalism, assessment of concepts in terms of epistemic values ensures that concepts are not formed arbitrarily. Thus, operational concepts of depression will proliferate only in so far as each demonstrably serves a useful function. Furthermore, the process of iterative revisions ensures that old operational concepts are replaced by new ones that fare better in terms of epistemic values. This

kind of revision already takes place: the process of constructing a measure in psychology is characterized by proposing test items (i.e. questions), testing their usefulness or validity, deleting poor items and re-testing the resulting revised measure (e.g. de Vet et al., 2011). So, the proliferation of operational concepts is not necessarily as wild as one might suspect.

What about comparability? If each measure of well-being defines its own concept, then it is possible that different scientists will arrive at very different, even contradictory generalizations on the basis of different measures. But this problem is mitigated by the fact that one of the criteria for a useful concept is similarity to a pre-theoretic or earlier scientific concept. In order to ensure conceptual continuity, and to avoid the objection that one is changing the subject, every operational concept of well-being should incorporate some pre-theoretic connotations of well-being. Indeed, this kind of conceptual continuity can be established by at least one already existing method of measure validation. To ensure so called face validity, researchers ask experts and laypeople to assess whether or not test questions pertain to the intended concept. For instance, when Tennant et al. validated a mental health and well-being questionnaire called the Affectometer 2, they asked focus groups about whether or not they thought the questions on the Affectometer 2 pertain to mental health and well-being (Tennant et al. 2006). While such a process does not ensure that different measures of mental well-being capture exactly the same aspects of mental well-being, it is likely to rule out gross misalignments in generalizations made on the basis of each measure.

Another way to build comparability between two operational concepts is to bypass the question of content and focus on statistical correspondences. For instance, Zimmerman et al. (2004) looked at how subjects responded to two depression measures, the Hamilton Depression Rating Scale (HDRS) and the Montgomery Åsberg Depression Rating Scale (MADRS), and then used linear regression to see how scores on MADRS correspond to scores on HDRS. In particular, Zimmerman et al. sought the MADRS total score that corresponds to HDRS scores equal or below 7, because in depression research this latter score is often treated as an indicator of remission from depression. Now, it is well-known (and easy to check) that HDRS and MADRS conceptualize depression in somewhat different ways. For example, HDRS lays much more emphasis on somatic symptoms than MADRS. The study of Zimmerman et al. can therefore be understood as an attempt to build comparability between MADRS-based claims about remission and HDRS-based claims about remission despite the fact that the two measures do not capture the same concept of depression.

By providing the above examples I do not mean to suggest that these methods have no downsides or blind spots. The point is just that researchers can build some types of comparability into their measures, even if they accept that the measures define different operational concepts. The problems of proliferation and comparability are not insurmountable.

5.2 Antirealism, caution, and the import of empirical evidence

Consider a fictional intelligence researcher in the early twentieth century. She has constructed a test that requires the test-taker to solve various problems: there are memory-related tasks, vocabulary tests, mathematical puzzles, tests of inter-personal skills, and so on. Suppose the intended long-term goal of the scientist is to measure intelligence understood in non-operational terms, but she does not know what kind of test-independent attribute or property determines success on these kinds of tests. In other words, she does not know whether test success is driven by a specific, unified cognitive skill, a cluster of such skills, a physical property of the brain, learned skills, test-wiseness, acculturation, or some other unobservable, but stable and easily characterizable property. What should she say when people ask what her measure measures? One option is to say that, for now, she defines intelligence in terms of the test. Intelligence is what the test tests, she argues, echoing Edwin G. Boring's (1923) famous dictum.

How is this a solution rather than a nasty cop-out? The answer is in the above phrase “for now”. Because the researcher is cautious, she does not want to hazard a guess, even a relatively educated one, at what kind of test-independent attribute causes or determines test responses—indeed, some of the early developers of intelligence tests were reluctant to guess what underlies differences between test takers scores on such tests (see e.g. quotes in Michell (2012)). In other words, since she does not understand the nature of what her measure captures, she cannot say whether it can be legitimately characterized as intelligence. But the researcher still thinks that the tests she has selected reflect pre-theoretic connotations of intelligence, that is, that there is continuity between the operational concept of intelligence and pre-theoretic connotations of intelligence. Indeed, intelligence connotes problem-solving, so a test involving problem-solving has a certain continuity with pre-theoretic views of intelligence. Furthermore, the researcher believes her measure has potential to be developed, and that an appropriately revised measure might well capture something that can legitimately be denoted with a non-operational concept of intelligence. For these reasons, the label “intelligence” is appropriate for the operational concept the test battery defines, even though the researcher does not know what kind of test-independent process or entity determines test responses.

In a situation such as the one outlined above, operationalism does not entail anti-realism. In other words, the act of defining the target attribute in terms of a test does not signal (i) a denial of the test-independent existence of intelligence, or (ii) a denial of the possibility of epistemic access to the determinants of test responses. Rather, this kind of operationalism is a temporary solution motivated by caution. What the caution-driven operationalist is implicitly saying is this: there might be a defensible conception of intelligence that has a test-independent referent, but I do not yet know what that concept would look like. Furthermore, that concept might be capturable

by empirical means, but I do not yet know what those means are. To avoid unwarranted inferences, I will go for an operational notion of intelligence.

Caution-driven operationalism utilizes (at least) two Carnapian principles: iteration and the idea that human judgment and pushback from reality are intertwined in the process of concept formation. To see this, consider how the revision of a concept of intelligence might go. The first revisions will likely be based on things like correlations between tests, stability of test results over time, and other such “operation-level” properties, that is, properties that do not in and of themselves say very much about what kind of test-independent attribute determines test results.⁹ But revision of the measure in light of evidence of stability and correlations between tests might help researchers zoom in on that determinant in the long run. This is because one explanation of why test results are stable over time is that the test captures the same test-independent attribute on each administration. Likewise, one explanation of why two tests correlate is that they measure the same attribute. If researchers are lucky, the triangulation of various types of evidence, and revision of the measure in light of that evidence, might produce knowledge about the nature of the determinant of test-responses (on triangulation, see e.g. Kuorikoski & Marchionni, 2016). Through these iterative steps, then, scientists might gradually gain enough confidence to make the step from a cautious operational concept to a bold non-operational concept.

It is not a novel idea that validation can help researchers zoom in on what exactly a test measures. My aim here is just to give an interpretation of these standard validation activities in terms of conceptual engineering, and to thereby formulate a version of operationalism that is not doomed to antirealism. Vis-à-vis antirealism, the point of the above (admittedly oversimplified) depiction of operationalist conceptual engineering is this. In this process, every conceptual revision—that is, every Carnapian epistemic iteration—whether it is from an operational concept to another operational one or from an operational concept to a non-operational one, involves a combination of human judgment and empirical evidence that testifies for reality. When doing the operational-to-operational-move, a researcher may well admit that test-independent properties exist and that they exert an influence on the way their measure behaves. But she chooses not to formulate her concept in terms of test-independent properties, because she judges that the evidence is not sufficient to warrant claims about what the relevant test-independent properties are like. Test-independent reality does push back, but the researcher chooses to not describe that reality when characterizing her target concept. This kind of caution-driven operationalism is perfectly compatible with realism and may indeed end up morphing into non-operational measurement in the long run.

These points about epistemic iteration are meant to illustrate that operationalists can accept the existence and possibility of epistemic access to test-independent entities, and hence need not be anti-realists. But do the tools psychologists currently use allow researchers to bootstrap their way from operational concepts to non-operational one? I do think that factor models, think aloud -studies, expert interviews, reliability analyses

⁹ Two tests might correlate e.g. because (i) they are biased in the same way, ii) they tap two distinct, causally related attributes, (iii) they tap two distinct, accidentally related attributes, or (iv) they track different aspect of a multidimensional attribute.

and other psychometric techniques allow scientists to gradually form a more coherent picture of the nature of the test-independent phenomena tests track. But this takes decades. For instance, in the last century, intelligence researchers' have made considerable progress regarding the nature of the entity or process that determines IQ test responses, but multiple characterizations of the nature of intelligence continue to be compatible with the existing evidence base (Van Der Maas et al., 2006). Luckily, I do not need to come up with an instant psychometric method for uncovering test-independent attributes here. In fact, the difficulty of working one's way from operational to non-operational concepts is good news for caution-driven operationalism. Because the epistemic road to credible non-operational concepts is so treacherous, it is good that researchers have a defensible operationalist position to fall back on, when the alternative is a very uncertain inference to a test-independent phenomenon.

5.3 Vetting values, respecting reality

The arbitrariness charge is now easy to rebut. I have already argued that conceptual engineering inspired operationalism uses epistemic values to assess proposed concepts. Concepts that fare poorly in this assessment are discarded, which means that it is not the case that anything goes. But doesn't the problem come back at the level of values? If we allow scientists to determine what properties a concept should have, and any property goes, then aren't researchers back to drowning in arbitrary concepts? I believe that scientific education, peer-review, criticism and public debate will make sure that not any value goes when it comes to concept formation. That is, the processes through which scientists justify their concepts to colleagues (and in some cases to the public) will tend to weed out poor valuative arguments for concepts. For instance, if a scientist discards the value of continuity (i.e. Carnap's similarity) when explicating well-being, other scientists and laypeople are likely to reject the resulting notion as irrelevant to the science of well-being.

What about Maul and McGrane's (2017, p. 2) argument that "when measurement is freed from any reality constraints, the inevitable outcome is [...] that anything [goes]"? Above I showed that operationalists need not deny the existence of, or the possibility of epistemic access to non-operationally characterized psychological qualities. I also argued that reality seeps in when researchers assess concepts in terms of values. That is to say, when scientists revise concepts in light of different epistemic values, they are often "consulting reality" whether or not they allow themselves to make claims about the nature of that reality. For instance, when researchers investigate statistical associations between two operational concepts of well-being, the real reasons that drive people to respond similarly or dissimilarly to each questionnaire will tend to show up in the association scientists observe (unless it is confounded, but of course there are techniques for handling that). Operationalism in the conceptual engineering mode is not freed from the constraints of reality, and therefore Maul and McGrane's argument does not apply.

5.4 Precision, accuracy, and steps in between

The final criticism I will deal with states that operationalism is suspicious, because it cannot make sense of two central aspects of scientific measurement: error and accuracy. Now, I think it is true that operationalists cannot, strictly speaking, evaluate the accuracy of their own measures, because they want to refrain from claims about what test-independent attribute their measure tracks. Put in another way, operationalists are always measuring their target concept by definition, and thus there is nothing they can be in error about. But even though operationalists cannot assess accuracy, they can evaluate precision—that is, the consistency of results when numerous measurements are taken under similar conditions. Luckily the psychometric theory of reliability is really a theory of precision or repeatability (Cronbach et al., 1963; Kline, 1998, p. 26; Nunnally & Bernstein, 1994, p. 248), which means that operationalists have their tools cut out for them.

More ambitiously, operationalists can take stabs at the assessment of accuracy as soon as they dare to take off their operationalist hats and propose non-operational versions of their operational concepts. Recall, again, that operationalists can admit that something test-independent underlies their measure—their scientific claims are just not about that something, at least not yet. With this admission, operationalists might occasionally be tempted to propose a non-operational revision of the operational concept, and then proceed to check for accuracy. For example, suppose an operationalist researcher hypothesizes that her well-being related questionnaire tracks differences in people's cognitive judgment of how life is going for them overall. One of the first steps in checking for accuracy is to ensure that people indeed engage in such a cognitive judgment when they answer the questionnaire. One way to assess this is to do a think-aloud -study, where subjects speak out loud their thought process when they complete the questionnaire (cf. Al-Janabi et al., 2013). If the thinking process matches the hypothesized process, researchers might then go on to investigate how accurately the measure tracks changes in that process. With mounting evidence, they might eventually be convinced that the questionnaire accurately tracks a non-operational conception of well-being, namely, differences in people's cognitive judgment of how life is going for them overall.

6 Conclusion and implications for psychology

I have argued that the perspective of conceptual engineering allows us to formulate an operationalism that is able to resist some of the main criticisms previously levelled against it. This operationalism is not motivated by antirealism but rather by caution. Its proponents revise their concepts iteratively in light of evidence of the fulfillment of (epistemic) values, which are vetted in scientific discourse. These iterations are also informed by the way a test-independent property or properties influences the behavior of the measure that is under construction and validation. But the operationalist, being operationalist, is unwilling to extend their inferences to the nature of test-independent properties when those properties are poorly understood. Still, the operationalist project

might turn into a non-operationalist project if at some point evidence warrants sufficient confidence in characterizing the target concept in non-operational terms.

Although I have not argued for it here, I think that this kind of operationalism is needed in psychology. The reason is two-fold. With all the bad press operationalism has received, a scientist who wants their work respected is better off being vague about what their measure measures instead of declaring operationalist commitments. Indeed, an oft-cited methodology article claims that “[b]eing precise does not make us operationalists”, which would be an unnecessary assurance, if vague definitions were not sometimes used to pre-empt accusations of operationalism (Wilkinson, 1999, p. 597). But vagueness leads to confusion. And as I have argued, there is debate about whether or not psychologists’ bread-and-butter validation methods reliably ensure that measures indeed track non-operational concepts. I think that a defensible form of operationalism might help us unravel this confusion, because it gives psychologists license to be open about the potential limits of their measures and concepts. My concrete proposal, then, is that researchers should openly admit to operationalism—or otherwise come up with an argument to support a richer, test-independent reading of the claims they make. The rhetorical move where operational conceptions magically give rise to non-operational claims does not serve psychology well.

Another, related reason for the need for operationalism is that in psychology it is incredibly hard to establish the claim that a measure tracks an acceptable, non-operational concept. The reasons for this are numerous, but consider the following, for example:

- People learn, remember, tire, and get bored, which makes it difficult to construct a measure that reliably tracks the same attribute from one administration to another.
- People react to testing, complicating the inference from the testing situation to “the real world”.
- People lie, exaggerate and know themselves poorly, which biases test responses.
- Different people might interpret questions differently, which threatens comparability.
- The processes underlying psychological phenomena, such as depression, are likely so complicated that it is hard to enumerate and take into account all the factors that are relevant to measurement.
- Many people have a stake in the battle for how psychological properties, such as intelligence, are conceptualized. Therefore, the adequacy of *any* proposed measure is likely to be questioned by *someone*.

Psychologists and psychometricians have of course developed an impressive array of tools to overcome these problems, but many concepts still resist measurement in non-operationalist terms. Operationalist concepts might in some cases be the best scientists can have. Those concepts can serve useful functions while also functioning as stepping stones in efforts to create measures of non-operational concepts. An operational conception of depression, for instance, can be useful for the prediction of remission and relapse. An operational concept of well-being can

tell researchers something about how a divorce or a promotion changes people's lives.

Naturally, while researchers are still building non-operational concepts, the usage of operational concepts comes with a threat of confusion of its own. If scientists use various conceptualizations of well-being, they might end up with differing views of how well-being is affected by divorce or a promotion. The tools of comparability building discussed in Sect. 5 mitigate this somewhat but might not resolve it entirely. The least scientists can do, I think, is that when they make scientific claims on the basis of a given operational concept, they clearly index the claim to the relevant measure. For example, a claim about the relation between income and well-being would take the form: "Well-being is associated with income in such and such a manner, when well-being is construed in terms of satisfaction of preferences as expressed in questionnaire Q." This may seem cumbersome, but similarly precise definitions are already considered methodological best practice in psychology (Wilkinson, 1999).

Acknowledgements I thank three anonymous reviewers for their helpful comments. I am also grateful to participants of the Concept Formation in the Natural and Social Sciences -workshop (University of Zurich, 2018) for conversations and comments.

Funding Open access funding provided by National Institute for Health and Welfare (THL).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alexandrova, A. (2016). Can the science of well-being be objective? *The British Journal for the Philosophy of Science*, 69(2), 421–445. <https://doi.org/10.1093/bjps/axw027>
- Alexandrova, A., & Haybron, D. (2016). Is construct validation valid? *Philosophy of Science*, 83(5), 1098–1109. <https://doi.org/10.1086/687941>
- Al-Janabi, H., Keeley, T., Mitchell, P., & Coast, J. (2013). Can capabilities be self-reported? A think aloud study. *Social Science & Medicine*, 87, 116–122.
- Boring, E. G. (1923). Intelligence as the tests test it. *New Republic*, 36, 35–37.
- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71(3), 425–440. <https://doi.org/10.1007/s11336-006-1447-6>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110(2), 203–219. <https://doi.org/10.1037/0033-295X.110.2.203>
- Bridgman, P. W. (1959). *The way things are*. Harvard University Press.
- Brun, G. (2016). Explication as a method of conceptual re-engineering. *Erkenntnis*, 81(6), 1211–1241. <https://doi.org/10.1007/s10670-015-9791-5>
- Cappelen, H. (2018). *Fixing language: An essay on conceptual engineering*. Oxford University Press.
- Carnap, R. (1934). On the character of philosophic problems. *Philosophy of Science*, 1(1), 5–18.
- Carnap, R. (1950a). Empiricism, semantics, and ontology. *Revue Internationale De Philosophie*, 4(11), 20–40.
- Carnap, R. (1950b). *Logical foundations of probability*. University of Chicago Press.
- Chang, H. (2004). *Inventing temperature: measurement and scientific progress*. Oxford University Press.

- Chang, H. (2009). Operationalism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, (Fall 2009).
- Chang, H. (2017). Operationalism: Old lessons and new challenges. In N. Mößner & A. Nordmann (Eds.), *Reasoning in Measurement* (pp. 25–38). Routledge.
- Creath, R. (Ed.). (1990). Introduction. In *Dear Carnap, Dear Van: The Quine-Carnap Correspondence and Related Work*. University of California Press.
- Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory†. *British Journal of Statistical Psychology*, 16(2), 137–163. <https://doi.org/10.1111/j.2044-8317.1963.tb00206.x>
- de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511996214>
- Dutilh Novaes, C. (2020). Carnapian explication and ameliorative analysis: A systematic comparison. *Synthese*, 197, 1011–1034. <https://doi.org/10.1007/s11229-018-1732-9>
- Evans, A., & Rooney, B. J. (2011). *Methods in psychological research* (2nd ed.). Sage.
- Feest, U. (2005). Operationism in psychology: What the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences*, 41(2), 131–149. <https://doi.org/10.1002/jhbs.20079>
- Feest, U. (2010). Concepts as tools in the experimental generation of knowledge in cognitive neuropsychology. *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 4(1), 173–190.
- Feest, U. (2012). Exploratory experiments, concept formation, and theory construction in psychology. In U. Feest & F. Steinle (Eds.), *Scientific concepts and investigative practice*. Berlin: De Gruyter.
- Flanagan, O. (1980). Skinnerian metaphysics and the problem of operationism. *Behaviorism*, 8(1), 1–13.
- Frank, P. G. (1956). *The validation of scientific theories*. Beacon Press.
- Gillies, D. A. (1972). Operationalism. *Synthese*, 25(1/2), 1–24.
- Green, C. D. (1992). Of immortal mythological beasts. *Theory & Psychology*, 2(3), 291–320. <https://doi.org/10.1177/0959354392023003>
- Hand, D. J. (2016). *Measurement: A very short introduction*. Oxford University Press.
- Haslanger, S. A. (2000). Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Noûs*, 34(1), 31–55.
- Hempel, C. G. (1954). A logical appraisal of operationism. *The Scientific Monthly*, 79, 215–220.
- Hempel, C. G. (1966). *Philosophy of natural science*. Prentice-Hall.
- Hull, D. L. (1968). The Operational Imperative: Sense and Nonsense in Operationism. *Systematic Zoology*, 17(4), 438. <https://doi.org/10.2307/2412042>
- Kline, P. (1998). *The new psychometrics: Science, psychology and measurement*. Routledge.
- Kuorikoski, J., & Marchionni, C. (2016). Evidential diversity and the triangulation of phenomena. *Philosophy of Science*, 83(2), 227–247. <https://doi.org/10.1086/684960>
- Leahey, T. H. (1980). The myth of operationism. *The Journal of Mind and Behavior*, 1(2), 127–143. <https://doi.org/10.2307/43852818>
- Lovett, B. J., & Hood, B. (2011). Realism and operationism in psychiatric diagnosis. *Philosophical Psychology*, 24(2), 207–222. <https://doi.org/10.1080/09515089.2011.558498>
- Markus, K., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.
- Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, 15(2), 51–69. <https://doi.org/10.1080/15366367.2017.1348108>
- Maul, A., & McGrane, J. (2017). As pragmatic as theft over honest toil: Disentangling pragmatism from operationalism. *Measurement: Interdisciplinary Research and Perspectives*, 15(1), 2–4. <https://doi.org/10.1080/15366367.2017.1342484>
- Maul, A., Wilson, M., & Torres Irribarra, D. (2013). On the conceptual foundations of psychological measurement. *Journal of Physics: Conference Series*, 459.
- McClimans, L., Browne, J., & Cano, S. (2017). Clinical outcome measurement: Models, theory, psychometrics and practice. *Studies in History and Philosophy of Science Part A*, 65–66, 67–73. <https://doi.org/10.1016/j.shpsa.2017.06.004>
- Meehl, P. E. (1995). Bootstraps taxometrics: Solving the classification problem in psychopathology. *American Psychologist*, 50(4), 266–275. <https://doi.org/10.1037/0003-066X.50.4.266>
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Lawrence Erlbaum Associates.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–383. <https://doi.org/10.1111/j.2044-8295.1997.tb02641.x>
- Michell, J. (2008). Is psychometrics pathological science. *Measurement: Interdisciplinary Research & Perspective*, 6(1–2), 7–24. <https://doi.org/10.1080/15366360802035489>

- Michell, J. (2012). Alfred binet and the concept of heterogeneous orders. *Frontiers in Psychology*, 3, 1–8. <https://doi.org/10.3389/fpsyg.2012.00261>
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. McGraw-Hill.
- Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, 12, 286–302.
- Salmon, W. C. (1994). Carnap, Hempel, and Reichenbach on Scientific Realism. In W. C. Salmon & G. Wolters (Eds.), *Logic, Language, and the Structure of Scientific Theories* (.). University of Pittsburgh Press.
- Schärp, K. (2013). *Replacing Truth*. Oxford University Press.
- Shepherd, J., & Justus, J. (2015). X-Phi and Carnapian explication. *Erkenntnis*, 80(2), 381–402.
- Stevens, S. S. (1935). The operational basis of psychology. *The American Journal of Psychology*, 47(2), 323–330.
- Strawson, P. F. (1963). Carnap's views on constructed systems versus natural languages in analytic philosophy. In P. A. Schlipp (Ed.), *The Philosophy of Rudolf Carnap*. La Salle, Ill: Open Court.
- Tennant, R., Fishwick, R., Platt, S., Joseph, S., & Stewart-Brown, S. (2006). Monitoring positive mental health in Scotland: Validating the Affectometer 2 scale and developing the Warwick-Edinburgh Mental Well-being Scale for the UK. Edinburgh: NHS Health Scotland. <http://www.healthscotland.scot/media/1719/5776-affectometer-wemwbs-final-report.pdf>.
- Tiberius, V. (2006). Well-being: Psychological research for philosophers. *Philosophy Compass*, 1(5), 493–505.
- Trout, J. D. (1998). *Measuring the intentional world : Realism, naturalism, and quantitative methods in the behavioral sciences*. Oxford University Press.
- Van Der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review*, 113(4), 842–861. <https://doi.org/10.1037/0033-295X.113.4.842>
- Vessonen, E. (2021). Respectful operationalism. *Theory & Psychology*, 31(1), 84–105. <https://doi.org/10.1177/0959354320945036>.
- Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2), 158–177. <https://doi.org/10.1037/h0074428>
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604.
- Zimmerman, M., Posternak, M. A., & Chelminski, I. (2004). Derivation of a Definition of remission on the Montgomery-Asberg depression rating scale corresponding to the definition of remission on the Hamilton rating scale for depression. *Journal of Psychiatric Research*, 38(6), 577–582.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.