European Journal for Philosophy of Science (2021) 11: 76
https://doi.org/10.1007/s13194-021-00365-6

PAPER IN GENERAL PHILOSOPHY OF SCIENCE

Representation in measurement

Elina Vessonen[1]

Received: 21 August 2020 / Accepted: 26 March 2021 / Published online: 28 July 2021

Abstract
The Representational Theory of Measurement (RTM) is the best known account of the kind of representation measurement requires. However, RTM has been challenged from various angles, with critics claiming e.g. that RTM fails to account for actual measurement practice and that it is ambiguous about the nature of measurable attributes. In this paper I use the critical literature on RTM to formulate Representation Minimalism – a characterization of what measurement-relevant representation requires at the minimum. I argue that Representation Minimalism avoids the main problems with RTM while acknowledging its usefulness as the formal foundation of representation in measurement.

1 Introduction

Measurement is representation – that much is usually agreed upon. The best known and most ambitious account of the kind of representation measurement requires is called the Representational Theory of Measurement (RTM). RTM received its authorative expression in three volumes entitled *Foundations of Measurement*, which were written by philosopher Patrick Suppes and psychologists R. Duncan Luce, Amos Tversky and David Krantz. The books appeared in 1971, 1989 and 1990. Since then, RTM has played a role in decision-theory in economics and philosophy, as well as in economic theory more generally.

Nonetheless, the content and usefulness of RTM are debated. Some argue that RTM is a formal theory specifying conditions of measurement, others argue that RTM is an epistemological theory specifying how to go about measuring (Tal, 2012 introduces alternative interpretations). Some argue that RTM is useful for measurement (Baccelli, 2020; Heilmann, 2015), others argue that it may be redundant in some contexts (Angner, 2011). Some argue that RTM may have formal merits but that it is useless for the practical execution of measurement (Mari et al., 2017; Reiss, 2008), still others say that RTM fails as an approach to measurement and should

✉ Elina Vessonen
elina.vessonen@gmail.com

[1] Visiting Researcher, Finnish Institute for Health and Welfare, Helsinki, Finland

be replaced by something else (Michell, 2005, 2020). Clearly, RTM is not the kind shared background one can lean on when debating and improving measurement practices.

The benefit of a rich, critical literature on RTM is that it signposts controversial aspects of representation. The literature on RTM therefore provides a roadmap of pitfalls to avoid when developing a new account of measurement-relevant representation. In this paper, I set out to define minimal conditions for representation in a measurement context. Accordingly, my account is called Representation Minimalism.

Why would one want minimal conditions of measurement-relevant representation, rather than a bold, full-fledged account? I think that neutral, common ground is sorely needed in the measurement literature. There are various on-going debates that are so multifaceted that it is hard to disentangle the exact source of the controversy. This increases the risk of people talking past each other. As an example, consider the literature on RTM. Some people argue that RTM is useful for measurement (Cartwright et al., 2016; Heilmann, 2015), others argue that it is untenable or useless (Michell, 1997, 2020). This looks like a genuine controversy (as opposed to a terminological one), and some authors treat it as such. But if different authors characterize RTM in different ways, the arguments of proponents of seemingly opposing views might in fact be compatible with each other. What looks like a substantial controversy, might not be one.

A similar situation is present in the literature on operationalism – operationalism stands for some version of the idea that it is permissible to define target attributes, e.g. depression or length, in terms of the measurement procedure intended to capture that concept. Many authors argue vehemently against operationalism (Borsboom et al., 2003; Maul & McGrane, 2017), while others find forms of it defensible (Chang, 2017; Feest, 2010). Is this a genuine controversy or just people talking about different things under the same term? By now operationalism has so many different definitions and characterizations, that it is hard to tell whether critics and proponents are talking about the same thing. Some common, definitional ground regarding the nature of successful measurement would help researchers name the exact sources of their disagreement – if in fact there is disagreement.

There is also considerable controversy about the measurability, and numerical representability, of individual concepts or constructs, especially in the human sciences.[1] Intelligence, well-being, health and countless other attributes are measured in a myriad of ways without any consensus on what it takes to adequately numerically represent them. These controversies have tangible effects: for example, unresolved debates about adequate measurement of depression lead to situations where a drug looks effective or ineffective depending on which measure is being used (Le Noury et al., 2015; Snaith, 1993). Another measurement-related debate is the one that asks whether there is some fundamental difference between the ways in which numbers represent attributes in the social sciences and in the natural sciences (Michell, 1986). A minimal account of measurement-relevant representation will not resolve these

[1] By "measuring a concept" I mean numerically representing the denotation of that concept, i.e. an attribute or a property of an entity.

debates. But recognizing a neutral common ground is a useful starting point. If we can find such a common ground, it is easier to articulate points of disagreement and to avoid situations where creators, users and critics of measures talk past each other.

After outlining Representation Minimalism (Sect. 2), I will use the literature on RTM to show that Representation Minimalism avoids many of the worries that scholars have had about the conception of representation that RTM *is taken to advocate*. I emphasised "*taken to advocate"*, because it is my view that RTM is often misunderstood by its critics. Indeed, I will argue that when RTM is appropriately interpreted, RTM and Representation Minimalism are allies, not alternatives (Sect. 3). Finally, I defend representation minimalism against some of the problems critics have diagnosed with RTM, i.e. its ambiguity about the ontology of empirical relations and its alleged incompatibility with measurement practices (Sects. 4 and 5, respectively). Section 6 concludes.

2 Representation minimalism

2.1 Scales in measurement practice

Consider an assignment of numerals to types of minerals:

Quartz ← 7.

Calcite ← 3.

The assignment consists of completely useless uninterpreted symbols unless we know what the symbols are meant to represent or be informative of. The user of these numerals wants to know (at least) two things before she allows herself to interpret the numerals as meaningful measurement results. First, what attribute do the numerals pertain to? Do they indicate the weight or surface temperature or hardness or something else entirely? Second: what measurement scale are we dealing with? This section will focus on this latter question. The notion of *scale* is a crucial tool for explicating measurement-relevant representation, which is why we need to familiarize ourselves with it.

Scales are part and parcel of the conceptual and statistical toolbox of measurement. Hence, scales are introduced in most textbooks and introductory courses on social scientific measurement (Embretson & Reise, 2000; Fiske, 1971; Howell, 2010; Kline, 1998; Lord & Novick, 1968; Nunnally & Bernstein, 1994; Osherson & Lane, 2018). But more familiar attributes from physical sciences, weight, temperature and hardness, are simpler cases for illustrating the most common measurement scale types, which were originally distinguished by psychophysicist S. S. Stevens (1946).[2] The following paragraph provides (what I take to be) an uncontroversial, broad brushstrokes account of how scale types are usually treated.

Our familiarity with weight measurement tells us that if the numbers indicate weight measured in grams, we can say, for example, that the ratio of the quartz sample's weight to that of the calcite sample is 2.3. In measurement jargon, we

[2] There are alternative classifications of scale types. See e.g. Velleman and Wilkinson (1993).

say that such comparisons are meaningful because weight is measured on a ratio scale. By contrast, we know from everyday usage of temperature scales that it is not sensible to say that the ratio of the temperature of the quartz sample to the temperature of the calcite sample is 2.3, when both are measured in centigrade. We could, however, compare the ratio of the *difference* between the temperatures of the two samples to the difference between the temperatures of some other two samples. In the language of measurement theory, we say that such comparisons are meaningful because the measurements are on an interval scale. Finally, if the numerals are measurements of mineral hardness on the less familiar Mohs hardness scale, the only thing the numerals are informative of is ordering. Quartz is harder than calcite, we can say, knowing that it has been assigned a higher value. But based on measurements on Mohs hardness scale, we know nothing about how much harder quartz is. That is because Mohs hardness scale for minerals is an ordinal scale.

These crude characterizations should be uncontroversial and acceptable to most scientists and philosophers who deal with measurement scales. It is also widely accepted that different scale types tend to allow for different arithmetic and statistical operations, although there is disagreement on how strict one should be regarding these rules (Borgatta & Bohrnstedt, 1980; Luce et al., 1990; Stevens, 1951; Velleman & Wilkinson, 1993). Many social scientists and statisticians agree that one cannot, for example, sensibly take the mean of ordinal values or meaningfully add ordinal values to each other. Another example: in empirical social sciences, undergraduates must learn by heart that ordinal variables require a Spearman correlation test while Pearson correlation test can only be applied to variables measured on an interval or a ratio scale. It must immediately be added that the "rules" for identifying permissible statistical tests based on scale type are constantly debated and even more often broken. But for now, it suffices to remember that scale types set some constraints on what arithmetic and statistical operations can be fruitfully applied to measurement results.

I think it is appropriate to say, based on the foregoing, that scales are typically treated as representational assumptions. For example, when they pertain to the attribute weight measured on a ratio scale, the numbers 7 and 3 (from the previous example) *represent* the empirically established fact that in some situation of interest the quartz sample is 4 g heavier than the calcite sample and that the ratio of the quartz sample's weight to that of the calcite sample is 2.3. The following examples from introductory resources enforce that a representational reading of scales is common:

> [When dealing with an ordinal measure of stress] we do not assume, for example, that the difference between 10 and 15 points **represents** the same difference in stress as the difference between 15 and 20 points. Distinctions of that sort must be left to interval scales. (Howell, 2010, p. 7 emphasis added).
>
> Interval scales are numerical scales in which intervals have the same interpretation throughout. As an example, consider the Fahrenheit scale of temperature. The difference between 30 degrees and 40 degrees **represents** the

same temperature difference as the difference between 80 degrees and 90 degrees. (Osherson & Lane, 2018, emphasis added)

The common reading of scales, then, is that claims about scale types are claims about representation – e.g. "Measure M represents attribute A on scale S". The next section capitalizes on this notion of scales to explicate the kind of representation measurement requires.

2.2 Representation minimalism – a definition of measurement-relevant representation

Consider the following definition:

Representation Minimalism (ReM). In measurement, a numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities, when entities are considered in terms of the target attribute.

As with many philosophical theories of scientific representation, this definition of representation capitalizes on the detection of structural similarity. The need for structural similarity is captured in the requirement that *relations in the representing numerical system mirror empirical relations between entities.* In other words, the relations that exist between numbers that are assigned to entities need to be similar to the relations that exist between those entities. For example, *ordering* of numerals should mirror *ordering* of entities, *equalities* of numbers should mirror *equalities* in the degree to which two entities possess a property, and so on, where the italicized words designate what structural aspects of the numerical system and the empirical system are similar to each other.

Of course, a general definition of representation in terms of similarity is notoriously elusive, not least because "everything is similar to everything else" *in some respects* (Isaac, 2013; cf. McLendon, 1955). To avoid a trivial definition of representation, one must say something about *the kinds of similarities* that are relevant for representation. Fortunately, in the measurement context, common scale types allow us to enumerate some similarity relations that are useful for representation without recourse to a general definition. The most common, measurement-relevant mirrorings are ones where: i) order relations between numbers mirror order relations between entities (ordinal scales), ii) (in)equalities of differences between numbers mirror (in)equalities of differences between entities (interval scales), and iii) (in)equalities of ratios of numbers mirror (in)equalities of ratios of entities (ratio scales) when entities are considered in terms of some attribute. These similarity relations are useful in virtue of being intuitive and recognizable to most people who are involved in measurement activities and who are therefore familiar with scales. There are other scale classifications and thus other interesting and potentially useful measurement-related mirrorings. For the purposes of the present discussion, the most common scale types suffice though.

The term "specified" is important in the definition: A numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities. Its role is to signal that, while different kinds of

similarity relations (e.g. ordering, inequality and equality of differences) can justifiably underwrite measurement-relevant representation, it is important to be explicit about the relations a particular numerical assignment mirrors. On their own, numbers lure us to apply familiar operations such as addition and calculation of averages. But in measurement, the results of these operations only have a meaning if the numbers have a specific mirroring relation to the entities that are of interest to the measurer. For example, the average of numerals that signal mere ordering has no empirical interpretation – if you assign numbers 1 to 10 to 10 children according to their relative heights and take the average of those numbers, the resulting number has nothing to do with the average height of the children. Hence, for the numerical representation to be appropriate, one must be clear about what similarity relations are being represented. The relevant relations need to be specified.

Consider now the part of the definition of ReM that reads: *when entities are considered in terms of the target attribute*. Recall from earlier that a purportedly measurement-relevant numerical assignment raises (at least) two questions: one about scale and one about the attribute the numerical assignment pertains to. If ReM neglected to mention attributes, its usage might lead to the following situation. Consider the following information about three individuals, A, B and C, and their orderings in terms of height and weight:

Height	$C \succ B \succ A \prec C$
Weight	$A \succ B \succ C \prec A$

With a definition lacking any mention of attributes, the following numerical assignment might be said to be an appropriate, measurement-relevant representation: $A \leftarrow 2$, $C \leftarrow 3$, $B \leftarrow 1$.[3] Such a numerical assignment would mirror the following order relations: C is taller than A and B, and A is heavier than B. But the representation is odd and difficult to interpret. We would typically not consider such a representation appropriate.

By mentioning the target attribute in the definition of ReM I have tried to fend off such interpretational difficulties. But what are target attributes and what is it for entities to be considered in terms of such an attribute?[4] I propose that at minimum, there must be a unified characterization of the target attribute that applies to all the represented relations between entities. For example, say someone has represented some relation between Tim, Tam and Tom with ordered numbers 3, 2, and 1. The demand for a unified characterization means that if the numbers 2 and 1 signal a difference in height, then all other pairings of numbers must also signal differences in height.

This condition is meant to eliminate the kind of uninterpretable hodgepodge representations that we saw in the previous example, where some numerical relations reflected height and others weight. The problem with such representations is that the user of the numerical representation cannot read the numerical relations in terms of the intended empirical relations, because any given pair of numerals could pertain to

[3] The numerals designate orderings, such that a higher number designates the heavier/taller individual and an equal number designates equal height/weight.

[4] For philosophical literature on properties, see Galluzzo and Loux (2015); Marmodoro and Yates (2016).

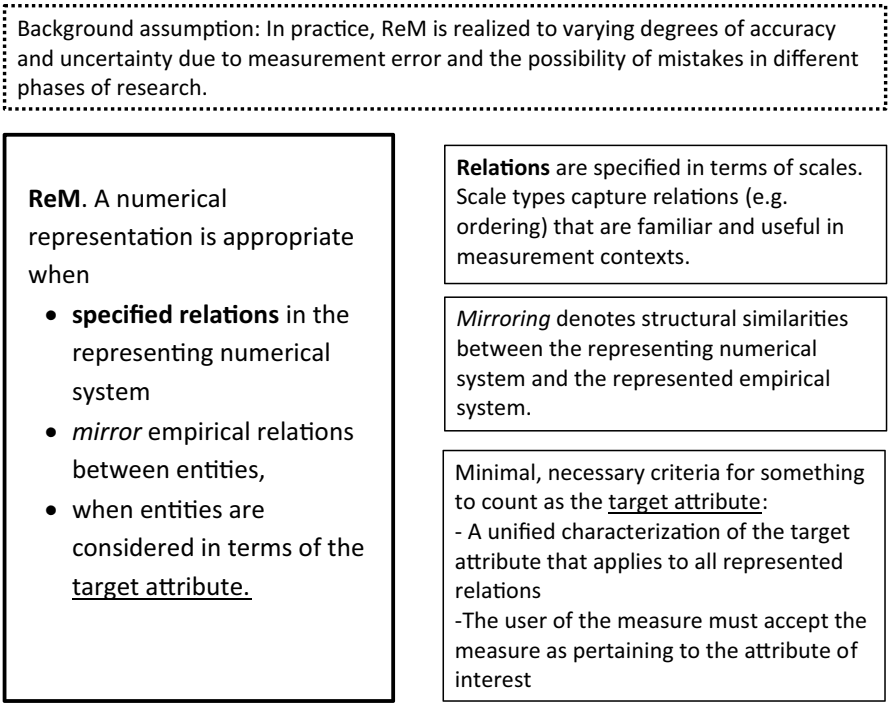

Fig. 1 A summary of Representation of Minimalism

weight or height. To avoid *this* kind of underdetermination, the target attribute must be characterized so that each pair of numerals represents the same attribute. This condition is still extremely permissive. Under it, acceptable characterizations would include e.g. height measured by eyeballing, total score on personality test T, the ratio of an individual's body temperature to the length of their eyebrow, and so on.

In addition, the target attribute must be the attribute or feature the user of the measurement instrument is *interested in* – hence the word "target". For example, it won't do to tell me to consider relations between entities using a ruler, if the attribute I am interested in is depression (and no advice is given on how to read the ruler in terms of depression rather than length). This is because as a measure-user, I am unlikely to accept the ruler as pertaining to the attribute I am interested in, i.e. depression.

These conditions for delineating the target attribute are human-dependant: it is up to us to formulate a unified characterization of an attribute and to decide whether the characterization captures the attribute of interest. The point of these conditions is not to imply that *attributes* are human-dependent, but rather to ensure the minimality of the resulting definition of representation. Formulated this way, we ensure that realists and anti-realists can buy in to the same definition of representation. Their disagreement will show up as a divergence in the kinds of things they accept as target attributes, as we will see in more detail in Sect. 4.

Taking ReM literally would mean that almost no current measurement procedure, whether in physics or psychology or anywhere in between, provides an appropriate representation. This is because ReM is silent about mistakes and measurement error. Due to random and systematic error, the numbers our measurement procedures yield never perfectly mirror empirical relations (unless they do so by definition but let us not get in to that here). We might also interpret evidence wrong and be mistaken about a hypothesized mirroring. To take these challenges into account, we would have to modify ReM to something like this:

ReM$_{\text{Evidence}}$. In measurement, we can tentatively accept a numerical representation as appropriate when we have evidence that specified relations in the representing numerical system mirror empirical relations between entities to an acceptable degree of accuracy, when entities are considered in terms of the target attribute.

Now, because this definition is a mouthful and not very user-friendly, I am going to categorize measurement error and mistakes as silent background assumptions. Thus, ReM should be read as a characterization of the kind of representation measurement aims for, but the attainment of which measurement error and mistakes prevent to varying degrees.

Figure 1 summarizes the resulting notion of measurement-relevant representation, explaining its key components. The definition is minimalist in many ways, and I have therefore called it Representation Minimalism (ReM). What I mean by minimalism is that the definition sets minimal constraints on the required representation – this is what measurement requires at minimum, although it might (be argued to) require a lot more in addition. First, ReM says nothing about the exact relations that need to be mirrored. What matters is that the relations are specified and consistently used, not that they are, for example, all and only the relations that interval and ratio scales are taken to represent. Second, ReM is minimalist regarding the kinds of attributes that are represented: it is not tied to a realist or any other metaphysical stance regarding attributes. Such minimalism does *not* mean that realists or proponents of other metaphysical views (concerning attributes) cannot commit to ReM. To the contrary, ReM sets minimal constraints for measurement-relevant representation, which the metaphysically-inclined may supplement as they see fit.

I believe that ReM is a necessary condition for measurement across the sciences, and the rest of this paper is dedicated to defending ReM as such. But ReM is likely not a sufficient condition for measurement. A full-blown account of measurement would likely specify how entities and the measurement instrument (causally) interact in order to produce an adequate representation (Cartwright et al., 2016). This paper does not attempt such a general account, although in Sect. 5 I will argue that ReM is compatible with common measure validation practices. Before that, a few words on the relationship between ReM and RTM.

3 ReM and RTM

3.1 Mathematics of scales

The Representational Theory of Measurement (RTM) is the most ambitious attempt to theorize measurement-relevant representation. One might think that by proposing a new definition of measurement-relevant representation, I am implying that RTM is a bad theory and that it should be replaced by the one I propose. But the way I – along many other measurement scholars – interpret RTM means that RTM and ReM are not two competing answers to the same philosophical question. In my view, RTM gives us formal tools for evaluating whether or not a given numerical representation of a given phenomenon is appropriate. ReM, on the other hand, is a more general definition of what measurement-relevant representation requires at minimum in any context.

In this section I am going to explicate the relationship between RTM and ReM in more detail. A thorough explication is needed for two reasons. Firstly, as the philosophical literature on measurement is growing, it is important to keep track of the relations between various theories. This prevents people from seeing controversy where there is none. This is especially important in this case, since I use the criticisms levelled against RTM to show that ReM does not fall prey to these objections, which might give the impression that RTM and ReM are rivals. Secondly, by outlining the way ReM and RTM can serve related but different functions, I underline that under certain interpretations, RTM too is a relevant theory of measurement. More specifically, I think that when RTM is interpreted as a formal theory of numerical representability, not as a theory of how to go about measuring, it too can serve useful functions without falling prey to objections commonly levelled against it. But before we get to the relations between ReM and RTM, a brief recap of the central aspects of RTM is in order.

According to the authors of *Foundations of Measurement*, where RTM received its canonical expression, measurement involves "the construction of homomorphisms from empirical relational structures of interest to numerical structures that are useful." (Krantz et al. 1971, 9). In mathematics, *homomorphisms* are many-to-one mappings from sets to other sets. In RTM, these mappings are established between specific types of sets, that is, the mappings relate *empirical relational structures* to *numerical (relational) structures*. Foundational measurement theorists commonly distinguish between four types of homomorphisms: ratio, interval, ordinal and nominal. As one can guess, types of homomorphisms are in fact what we have previously called scales.

I have already said that informally, differences between scales can be thought of as differences in what information the numbers (numerals) represent about the targets of measurement. But formally, different scales i.e. homomorphisms are characterized by the types of transformations they allow, that is, what are the rules for changing the numbers in a given numerical assignment. Ordinal scales, such as Mohs hardness scale for minerals, allow monotonic increasing transformations of the form $\phi \rightarrow f(\phi)$, where f is "a strictly increasing real-valued function of a real

variable" (Krantz et al. 1971, 11). These transformations are permissible, because a rule-bindingly transformed numerical assignment continues to represent the same empirical relational system as the original numerical assignment.[5] For example, if we are dealing with the ordinal Mohs hardness scale, the numerals in the numerical assignment.

Quartz $\leftarrow 7$,

Calcite $\leftarrow 3$

could be transformed by adding 2 to both numbers or multiplying them by 2, and the resulting numerical assignments would continue to mirror the structure ordinal scales are informative of, namely, ordering relations between entities (i.e. quartz would continue to be assigned a higher value which corresponds to the fact that it is harder). For interval scales, e.g. temperature measured in degrees Celsius or degrees Fahrenheit, the permissible transformations are of the form $\phi \rightarrow \alpha\phi + b$, $\alpha > 0$. Ratio scales, such as length and weight, allow for multiplicative transformation of the form $\phi \rightarrow \alpha\phi$, $\alpha > 0$. These latter two scales are also called quantitative scales.

So far, I have largely re-described the content of the previous section in the more technical and rigorous language of RTM. The core idea of RTM that we have not yet explored is that in order for relations between entities to be measured on a specific scale, those empirical relations have to fulfil certain constraints. RTM states these constraints in axioms. For example, the axioms that pertain to an ordinal scale are:

Let A be a finite set of objects, and $\succcurlyeq$ a binary relation on A. The relational structure ($\succcurlyeq$,A) can be meaningfully represented on an ordinal scale, iff for all a, b, c$\in$A,

1. Connectedness: Either a $\succcurlyeq$ b or b $\succcurlyeq$ a, and
2. Transitivity: If a $\succcurlyeq$ b and b $\succcurlyeq$ c, then a $\succcurlyeq$ c.

For example: the set A of objects denotes a set of minerals, and the relation $\succcurlyeq$ denotes a relation in terms of hardness, i.e. a $\succcurlyeq$ b is interpreted as "a is at least as hard as b".

Where do such constraints come from? The above claim about axioms of ordinal measurement, just like other axiomatizations in RTM, are backed up by theorems. In fact, proofs of two types of theorems fill the pages of *Foundations of Measurement* and other publications in the RTM tradition. A representation theorem establishes that if (sometimes *if and only if*) a given empirical relational structure of interest satisfies certain (non-contradictory) axioms, such as the ones described above, then a homomorphism ϕ to a certain numerical structure can be established. A uniqueness theorem establishes the permissible transformations of ϕ that also yield a homomorphism to the same numerical structure. In other words, the uniqueness theorem determines the scale type of the numerical assignment.

For example: if the hardness relation $\succcurlyeq$ satisfies connectedness and transitivity, then one can prove a representation theorem: there is a function ϕ from A to the

[5] Those versed in the technical literature might find the terms "transformation rules" and "meaningfulness" helpful in this context. I shall not introduce the terms here because I want to give an accessible gloss of RTM.

set of real numbers such that for all minerals a and b in A, $a \succcurlyeq b$ iff $\phi(a) \geq \phi(b)$. In informal terms, what is proven is that the hardness relation $\succcurlyeq$ holds between a and b if and only if the number associated with a is greater than or equal to the number associated to b. Another function ϕ' has the same property and thus constitutes a homomorphism to the same numerical structure as ϕ iff there is a strictly increasing function f such that for all a in A, $\phi'(a) = f[\phi(a)]$. In informal terms, in this case ϕ' is a permissible transformation of ϕ as long as it preserves the order of the numbers assigned to the objects. The upshot of the proofs is that mineral hardness can be represented on an ordinal scale.

3.2 The role of formal theory in representation

As noted in the introduction, multiple controversies surround RTM. Some contributors, for example, seem to think that RTM provides an account of how to go about measuring in practice, e.g. how to validate a measure or how to confirm scale type assumptions. More concretely, some scholars argue that RTM requires *direct observations* of relations that *perfectly* satisfy the relevant axioms (Angner, 2011; Borsboom, 2005; Mari, 2005; e.g. Michell, 1986).[6] From the observation that most successful measurement practices make no reference to direct observations or the fulfilment of axioms, it is then concluded that RTM is simply wrong about what successful measurement requires.

I think this critique of RTM does not hold water. This is because the critics have an incorrect, or at least an unfruitful interpretation of RTM. In my view, RTM is a formal theory of numerical representability: it tells us about the axiomatic conditions that guarantee the appropriateness of a certain numerical representation. RTM is silent about how to go about establishing that these conditions hold in a particular context, not because it is a bad theory, but simply because it was not meant to answer that question. Since various authors have defended this position elsewhere, I shall not dwell on the matter here (Baccelli, 2020; Decoene et al., 1995; Heilmann, 2015; Narens & Luce, 1993).

If RTM is treated as an account of the formal requirements of numerical representability, then RTM and Representation Minimalism are closely connected. In fact, RTM provides rigorous characterizations of the kinds of mirroring relations ReM requires. ReM relies on scale types to enumerate measurement-relevant mirrorings. RTM, by contrast, provides the conditions under which representations on a certain scale type are possible. That is, with its representation and uniqueness theorems RTM shows what kinds of empirical systems can be represented with a structure-informative numerical system, that continues to be informative of relevant structures under specific transformations on the numerical system. In short, RTM is the formal, definitional grounding of scale type assumptions.

[6] Borsboom (2005, ch. 4) eventually ends up characterizing RTM as a rational reconstruction of the measurement process, therefore departing from the observation-tied interpretation he first assigns to RTM.

To illustrate RTM's foundational role in characterizing scales, consider the transitivity condition, which RTM sets as a requirement for a structure to be represented on an ordinal scale. As the biconditional formulation of the axiom indicates, it is simply not possible to assign order-informative numbers to relations that violate the transitivity axiom. For example, consider a situation in which, for whatever way we measure and conceive of preference, Maya has strict preferences for Sacher cake over Pavlova, Pavlova over Black Forest and Black Forest over Sacher. It is evident that every assignment of numerals to cakes that attempts to capture that system of (strict!) order relations will fail, that is, at least one of the preference relations is not mirrored by the ordering of the assigned numerals. Our colloquial interpretation of ordinal scales as informative of order would not get off the ground without transitivity. This is how RTM provides the foundations of scale type assumptions.

Similarly, claims about interval or ratio measurement imply that a set of axioms holds, where those axioms can be proven to be jointly sufficient for constructing the relevant homomorphism. Notice though that for interval and ratio scale measurement, various axiomatizations are possible. In other words, there can be meaningful interval (ratio) level measurement of two attributes, even though the empirical relational systems they correspond to are different in some respects.

We can summarize the connection between RTM and ReM in the following three observations:

1. Representation Minimalism is defined in terms of mirrorings.
2. Scale types denote different kinds of mirrorings.
3. RTM provides formal foundations of scales.

From these conceptual links between relational structures, scales and mirrorings, we see that RTM provides formal foundations of ReM. The relevance of this ReM-RTM connection is that RTM-style axiomatizations can be used to evaluate the fulfilment of the requirements of Representation Minimalism. For example, if we are looking for evidence regarding representation of order relations, RTM gives us the transitivity and completeness axioms as empirical criteria. If there is empirical evidence for the fulfilment of transitivity and completeness, then there is evidence for the fulfilment of requirements of Representation Minimalism (in so far as the *specified relations* mentioned in ReM are indeed order relations).

This rendition of the relationship between ReM and RTM might give rise to the question: why do we need ReM if a particular interpretation of RTM already gives us the idea of numerical representability *and* a thorough system of theorems establishing the conditions of appropriate numerical representation? One way to think of ReM is that it is an extension of a particular interpretation of RTM, i.e. that ReM supplements RTM by engaging with questions that RTM has thus far not dealt with. These questions have to do with the role of error in establishing numerical representation (discussed in Sect. 2 above), the nature of attributes and empirical relations (Sect. 2 and 4) and the role of laws and models in validating numerical representations (Sect. 5 below). While I think this would be an acceptable framing, due to the controversies surrounding the nature of RTM I prefer to think of ReM as an independent account of representation

rather than as an extension of RTM. In other words, to avoid getting in to debates about what the founders of RTM meant by this-or-that phrase, I prefer to treat ReM as an account of representation that is compatible with certain interpretations of RTM, whether or not those interpretations where the ones intended by the authors of *Foundations of Measurement*.

Having now carved out the relationship between RTM and Representation Minimalism, we can move on to other concerns. RTM has evoked at least two kinds of criticisms. On the one hand, there is the worry that RTM fails to make sense of actual, successful measurement practices (e.g. Mari et al., 2017; Reiss, 2008). On the other hand, the exact nature of the concept "empirical relation" is ambiguous. These worries can be easily rewritten as challenges to Representation Minimalism. The rest of this paper will outline how Representation Minimalism deals with these two worries.

4 What are empirical relations?

One of the controversies surrounding RTM has to do with the exact nature of empirical relations. Are empirical relations real, measurement-independent patterns in the world that are merely revealed by measurement instruments? Or is the act of measuring in some sense constitutive of the relevant empirical relations? Should empirical relations be directly observable or can they be inferred from observations? (e.g. Angner, 2013; Vessonen, 2020) Representation Minimalism is non-committal in this regard. It only says that the relevant empirical relations have to be expressed in terms of an attribute that is of interest to the measure user and the relations have to have a unified characterization. The choice between types of empirical relations is determined by how the attribute of interest is conceptualized by the intended user of the measure. This way of delineating relevant relations is rare – in fact, I have not seen it laid out in this way anywhere else. In this section I will introduce some common readings of permissible empirical relations and relate them to ReM. In the minimalist spirit, I argue that ReM is neutral with respect to different ways of specifying empirical relations.

Consider, as an example, the Apgar score that quickly summarizes aspects of the health of an infant. At one minute and five minutes after delivery, the midwife or the doctor assesses five easily identifiable characteristics of the baby – heart rate, respiratory effort, muscle tone, reflex irritability, and colour – assigning a value of 0 to 2 to each characteristic and summing up the scores. Total scores of 7–10 denote good condition, while scores under 7 are thought to indicate that the baby might need urgent medical attention. This prediction is based on prior observations of the frequency with which infants that fall within a specific range of scores suffer from some serious health problems (such as brain damage). For example, studies show that the incidence of neonatal death is much higher among infants that are assigned scores 0 to 3 as compared to infants who are assigned scores 7 to 10 (e.g. Casey et al., 2001).

Does the numerical assignment Apgar score yields fulfil the requirement of measurement-relevant representation? The answer depends on what you take permissible empirical relations to be in the measurement context. In what follows I will go over some alternatives as to what Apgar score might be taken to represent, and problems with each alternative. I am *not* arguing that this or that interpretation is the correct way to read the Apgar score. I use the Apgar score merely to illustrate what kinds of relations numerical assignments might be taken to represent. I then show that my notion of representation, Representation Minimalism, is neutral with respect to these alternatives.

One might argue that the Apgar score represents empirical relations in terms of the attribute "what numbers get assigned when following the rules of the Apgar scoring system". For example, if baby Mia is assigned a "6" and baby Aija is assigned an "8", this represents the empirical relation that Mia received a lower score than Aija when they are assessed in terms of the rules of the Apgar scoring system. We might call this the *operationalist reading of empirical relations:* the numerical relational system represents the test procedure – the operation! – that yielded those numbers. One of the main problems with this approach is that measure-users are typically not interested in the procedure in and of itself, but only as far as it indicates some underlying, independent or "real" attribute of interest.

To contrast the operationalist reading, one might insist on (some version of) *realism about empirical relations*.[7] On this approach, it is not enough for the numerical structure to represent the procedure that produced the numbers. Instead, the numerical structure should represent empirical relations that exist independent of the testing procedure. For example, according to the realist, the Apgar score represents (or should represent) relations between infants in terms of health, illness or well-being, where health, illness and well-being are something that exist independent of the testing procedure. The realist would typically insist that what needs to be represented is relations that bring about or cause the testing procedure to yield the numbers it yields. For example, the test-independent health status of Mia and Aija is what causes the test procedure to assign an "8" to Aija and a "6" to Mia. The main problem with the realist reading is that it is difficult to determine whether the Apgar score tracks a plausible conception of health, illness or well-being.

We have seen the realist and the operationalist interpretation of measurement-related empirical relations. Are there others? In between realism and operationalism, one might insist, is another position: the relations that the numerical structure represents are the real, observed properties, such as skin colour and muscle tone, that the doctor or midwife reports with the Apgar score. Call this the *phenomenological reading* of empirical relations[8]: the numerical structure represents relations

[7] Wolff (2019), for instance, provides a compelling case for perspectival realism regarding measurement-related representation.

[8] The difference between the operationalist and the phenomenological reading is subtle. The operationalist says that the Apgar score represents relations between infants, when infants are compared in terms of the rules of the Apgar scoring system. The phenomonologist says that the test score represents relations between infants in terms of skin color, muscle tone and so on. On the phenomenological reading, we should be able to say how two infants differ in observed muscle tone when we see that they have been assigned scores 8 and 6, respectively. But clearly we are not justified in saying that. In the operationalist approach, we are merely saying that two infants differ in terms of their Apgar scores, which is obviously true.

in terms of immediate observations, which get reported via the Apgar scoring system.[9] The main reason I will not consider the phenomenological reading here is that it does not really count as a representation, that is, relations in the numerical structure cannot be interpreted or read in terms of relations in the empirical system. For example, the numbers that are assigned to Mia and Aija are not readily interpretable in terms of immediate observations, even though those numbers were assigned based on observations. The reason is that the difference in Aija's and Mia's respective scores could be due to a difference in observed muscle tone, or skin colour, or a combination of these, or a great many other combinations of observed properties. Similarly, if Mia and Aija received the same score "6", that would not be interpretable as representing the fact that Mia and Aija have a similar status in terms of observed properties. If the numerical assignment does not represent observed properties of the children, but is merely a product of considering those observations, it is not apt to consider this an approach to *representation*.

Another suggestion might be that alongside the operationalist and the realist interpretations, we should consider a predictive reading of empirical relations. For example, the Apgar score allows doctors and midwives to predict which infants are likely in need of urgent medical attention. The prediction is done based on experience and empirical data on the incidence of neonatal death, brain damage and other conditions in cohorts of babies that have been assigned a specific score. On these grounds, one might propose a *predictive reading* of empirical relations: what the numerical structure represents is (probabilistic) relations between infants in terms of their likelihood of suffering brain damage or another serious medical problem.

I think this reading can have useful functions but should not be considered an alternative to the realist and the operationalist readings. This is because the predictive reading always piggybacks on either the operationalist or the realist reading, depending on what kinds of empirical relations are being predicted. In the Apgar score case, what is predicted is a test-independent attribute such as brain damage or death. In other cases, operationally characterized attributes might be predicted, for example, if one uses Facebook posts to predict scores on a depression test (and that depression test is interpreted as representing test-dependent, i.e. operational relations). Under the predictive reading, then, a numerical structure is a representation of the likelihood of the occurrence of given realistically or operationally characterized empirical relations. The adequacy of a particular predictive reading therefore largely depends on the adequacy of the relevant realistic or operational reading, which is why I shall not consider predictive reading of empirical relations as an approach of its own.[10]

The operationalists and the realists frequently argue about the relative supremacy of each position (Lovett & Hood, 2011; Maul et al., 2016; on the opposition in

[9] Borsboom's constructivist reading of RTM (2005) could be thought of as similar to what I call the phenomenological reading here. As is evident from discussions above, I deny that RTM is committed to this reading of empirical relations.

[10] Here I dealt with the idea that a numerical structure may be taken to represent predicted operationalist or realist relations. There is, of course, also the case where a numerical assignment represents some operationalist or realist relations and is simultaneously predictive of other operationalist or realist relations. I mention this distinction here just to clarify the various alternatives one might consider.

psychometrics, see e.g. Michell, 2008). Some debates about the appropriate interpretation of RTM can be recast as debates about whether RTM is a realist or an operationalist approach. On the one hand, the fact that Krantz et al. (ch. 1 in 1971) used length as the prime example in their explication of RTM suggests the operational reading (or phenomenological reading),[11] on the other hand the three volumes of *Foundations of measurement* are littered with critical remarks about operationalism (especially chapter 1 of Volume 1 and chapter 22 of Volume 3).

The beauty of Representation Minimalism is that it does not hinge on a commitment to any of these positions. This is a virtue, because the point of a minimal account is to offer common ground for debates. Representation Minimalism takes both the operationalist approach and the realist approach on board by characterizing measurement-relevant representation in terms of an attribute that is of interest to the measure-user. If the measure-user is a realist, the representational capacities of a measure will impress them only if the measure yields a numerical structure that represents "realistically" characterized empirical relations. If the measure-user is an operationalist, the representational capacities of a measure will impress them only if the numerical structure represents operational empirical relations.

5 Models, minimalism and validation practice

RTM has been criticized for its failure to make sense of actual, successful measurement practices (e.g. Mari et al., 2017; Reiss, 2008). Critics note that when researchers create, validate and use measures, they make no mention of axioms, which is why the mathematical focus of RTM seems to fit poorly with actual measurement practice. A similar objection could be carved against ReM: when researchers successfully validate a measure, in particular their quantitative scale type assumptions, they make no reference to mirrorings. Hence characterizing measurement-relevant representation in terms of mirrorings must be wrong-headed. In this section I want to show why claiming that ReM is a necessary condition of measurement is compatible with the observation that successful measure validation practices indeed make no mention of mirrorings or mappings, let alone axioms.

How does one confirm scale type assumptions in practice? We can begin to look for answers in historical accounts of successful measure validation. Those histories rarely feature axioms. Rather, they involve theorizing and empirically testing law-like associations between attributes. A law-like association here means, roughly, a mathematically expressible, reasonably stable relation between the target attribute (A) and other attributes of interest (I): $A = f(I_1, I_2, I_3, \ldots I_n)$. In psychology and social sciences, such a mathematical expression of a stable relationship between target and other (indicator) attributes is called a *measurement model* (see e.g. Embretson & Reise, 2000). I will use this conception of a measurement model to explicate two types of approaches to

[11] The founder of operationalism (or at least its most famous champion) Percy Bridgman also used length as the showcase of operational analysis (Chang, 2009). The fact that length is a key example for both the operationalists and the RTM proponents may invite the interpretation that the two are the same or very similar approaches.

the confirmation of scale type assumptions: the direct axiom-based approach (DAB) and the indirect measurement model-based approach (IMM). I illustrate the two ways of confirming scale type assumptions with reference to temperature measurement.[12] Finally, I connect DAB and IMM to Representation Minimalism, showing that both approaches to scale type confirmation are compatible with Representation Minimalism. This way, I show that even though mirrorings are typically not *mentioned* in measurement practice, the practice of scale type confirmation still implicitly aims for and achieves the kind of representation Representation Minimalism requires.

When the developers of temperature measurement had achieved ordinal measurement on so-called thermoscopes, their ambitions to improve towards quantitative measurement did not direct them to axiomatic measurement theory. Rather, the focus was on the mathematical form of the law that governs the association between temperature and other relevant attributes, most importantly the association between temperature and the volume of an indicator substance, such as mercury, that fills the tube of a thermometer-to-be (under specified auxiliary conditions, e.g. pressure) (Chang, 2004, ch. 2). The hypothesized model, a linear relationship between temperature and the volume of the indicator substance, achieved support gradually and abductively through a combination of theorizing and experimenting (the details of this process do not concern us here). The theorizing and experimentation eventually justified the inference that temperature could be measured on an interval scale with an instrument that is filled with the relevant indicator substance (when the instrument's cross-sectional area is constant).

Why do (approximate) confirmations of such measurement models afford inferences to measurability on a quantitative scale? Superficially, a confirmation of the linear relationship between two *numerical* systems (sets of numbers), one called "volume" and one called "temperature", immediately ensures that the possible numerical assignments to levels of temperature are all linearly related to each other – it could not be otherwise if the linearity of the relation between numerical systems associated with temperature and volume is established. But to comply with our full characterization of scale type assumptions, we need something more substantial than the relation between two numerical systems: we need mappings from the permissible numerical systems to empirical relations. What does the confirmation of the measurement model have to be like in order for it to allow inferences of numerical-empirical mappings? In the case of temperature, what allows us to say that differences between numbers assigned map onto differences between entities in terms of temperature?

Notice first that in the case of temperature, the indicator attribute volume is known to be quantitative. Why? That volume is quantitative may seem trivial, and in fact people have treated volume as a quantitative attribute for millennia, because it is so intuitive to do so. But I will probe the foundations of those intuitions a little further here, since understanding the quantitative nature of volume will help me explain why confirmation of (some) measurement models affords inferences to scale types.

[12] The distinction between ways of confirming scale type assumptions is explicated here for the first time, but I rely heavily on Chang's (2004) account of the historical development of temperature measurement to explicate the distinction.

Loosely speaking, we know that the attribute volume is quantitative because we can readily observe that its subsystems are additive. That is to say, if we divide an entity into parts, the volume of the whole entity is always the sum of the volumes of the parts of the entity. (Such attributes are sometimes called "extensive" as opposed to "intensive" in the measurement literature.) Knowing this, we can mentally map the arithmetical operation of addition onto additive empirical operations involving the attribute volume. And recall that addition is a meaningful operation only for quantitative scales.

Such a chain of inference is, of course, a coarse-grained justification for the possibility of quantification. But it resonates well with mathematical measurement theories and exemplifies the core of more rigorous empirical approaches that proceed directly from RTM-style axioms. In the coarse-grained justification, one compares entities in terms of direct observations of the target attribute (e.g. how the volume of parts relates to the volume of the whole) and maps the observed relations to relations and operations on numbers (e.g. how two numbers relate to their sum) via mental association. A parallel but more rigorous empirical justification of scale type assumptions starts from measurement theory, which, as we have seen, maps uninterpreted relational systems to numerical systems via mathematical proofs. An empirical researcher can then capitalize on those proofs in her inference to scale type, if she manages to operationalize the axiomatic constraints in terms of the attribute of interest, and if her observations confirm that the constraints are fulfilled, at least approximately.

With this in mind, we can come back to inferences from measurement models to scales. To understand that inference, I think it is helpful to recast the confirmation of the linear relation between temperature and the volume of a thermometric substance in the following terms (recall, we are thinking about situations where the aim is to go beyond the superficial numerical-system-to-numerical-system association). What is happening, in effect, is a mapping of the known equalities and inequalities of differences in volume with the previously unknown equalities and inequalities in differences in temperature. In other words, we can think of the gradual confirmation of a linear law-like relation between temperature and volume of a given thermometric substance as the search for the conditions under which equal differences in volume map onto equal differences in temperature, and unequal differences in volume map onto unequal differences in temperature. This may sound like an unnecessarily complex way of putting the simple idea of confirming a linear law-like relation. But put this way it is easy to understand why the confirmation of the measurement model allows inferences to scale type. In the case of confirming the quantitative nature of volume, the procedure involved mapping arithmetic operations directly on empirical relations. Here, by contrast, the procedure starts from the mapping of empirical relations pertaining to volume to empirical relations pertaining to temperature,

and only then proceeds to map a numerical system onto the empirical relations thus discovered.[13]

The general point of this admittedly long-winded explanation is to motivate a distinction between two methods of establishing scale types in practice: the direct axiom-based approach (DAB) and the indirect measurement model-based approach (IMM). DAB proceeds from the mathematical measurement theory to empirical reality by taking the axioms (or something like them) at face value and finding direct (that is, simple and trivially justified) observational means of checking whether the axioms hold when entities are compared in terms of the target attribute. The establishment of the quantitative nature of volume is an example. In IMM, by contrast, one specifies a measurement model of the target attribute and its relations to other attributes and infers the scale type of the target attribute via confirmation of the relationship postulated in the model. The establishment of the interval measurement of temperature is an example. The division between DAB and IMM has grey areas, because of the blurriness of the dividing line between direct, simple observation on the one hand and indirect, inferential means of hypothesis confirmation on the other. There is no reason to let such greyness and blurriness alarm us.

It is not a novel idea that measure validation can involve something else than direct observations of the fulfilment of axioms. Nor is it new to claim that measures can be validated via modelling. For example, Tal (2012, 2016) has defined and defended what he calls the model-based approach to measurement with respect to time measurement, while McClimans et al. (2017) apply Tal's framework to argue that model-based considerations are central to measure validation in psychology. Much earlier, in 1934, psychologist Junius Flagg Brown wrote an article in *Erkenntnis*, where he analysed the validation of measures of e.g. weight, electrical potential and temperature in terms of the confirmation of (what in present terminology could be called) measurement models (because they express law-like tendencies mathematically) (Brown, 1934). If the claim about models in measure validation has been made some 80 years ago, what has been the point of this exercise?

My aim here is to use the DAB-IMM distinction to show that seemingly different practical processes of scale type confirmation share the same conceptual framework of measurement-relevant representation. I have explicated DAB and IMM in terms of volume and temperature to show that in both cases the underlying justification for the confirmation of the relevant scale type assumption has to do with the mirroring of empirical and numerical relational systems. In the case

[13] In this example, the mapping capitalizes on the *known* quantitative properties of volume – the confirmation of the linear relation amounts, in a sense, to an extrapolation from the known quantitative properties of volume to those of temperature. However, it is *not* a general feature of model-based confirmation of scale types that one attribute must be known to be quantitative prior to the confirmation of the relevant measurement model. The confirmation of a Rasch model is an example of a model-based confirmation that does not require prior quantitative information (Vessonen, 2020). I think this detail is not essential for the IMM/DAB distinction to do work for us, which is why I shall omit extensive discussion of the point here. The point of the IMM/DAB distinction is that seemingly different kinds of confirmation activities share in the same framework of representation.

of volume, the establishment of the mirroring is direct, consisting of the mental association of certain mathematical operations with observed relations in terms of volume. In the case of temperature, the establishment of the mirroring is indirect, proceeding from the empirical-to-empirical mirroring of differences in the volume of a thermometric substance to differences in temperature, and from there to the numerical-to-empirical mirroring pertaining to the interval scale assumption regarding temperature. The mirroring-based notion of representation grounds both cases, even though one is an example of DAB and the other an example of IMM. In this sense DAB and IMM share the same characterization of measurement-relevant representation, even though on the surface they may seem like very different kinds of activities. The upshot is that Representation Minimalism is compatible with common, model-based and non-model-based approaches to confirming scale type assumptions.

6 Representation in measurement

In this paper I have defined a notion of measurement-relevant representation. The definition I provided is:

Representation Minimalism (ReM). In measurement, a numerical representation is appropriate when specified relations in the representing numerical system mirror empirical relations between entities, when entities are considered in terms of the target attribute.

The rest of this paper defended this definition. First, I showed how ReM relates to RTM, arguing that RTM provides the formal foundations of ReM. Second, I showed that ReM is compatible with several common notions of the kinds of empirical relations measurement aims to represent. Third, and finally, I argued that ReM is compatible with plausible accounts of how scale types get confirmed.

As in almost any philosophical field, there is considerable controversy in the measurement literature. Often it is not clear what exactly the source of disagreement is. This is evidenced, for instance, by the confusion surrounding RTM: Is it a how-to account or merely the formal foundations of measurement? Are its axioms necessary for measurement or simply one approach among many? Is RTM an operationalist or a realist take on measurement? The motivation for formulating a minimal, bare bones account of representation is to try to find some common ground for these debates. If we can agree on some aspects of measurement-relevant representation, it is easier to keep track of the exact points of disagreement and to avoid talking past each other.

Acknowledgments I thank Anna Alexandrova, Denny Borsboom, Hasok Chang, Tim Lewens and two anonymous referees for their helpful comments on earlier drafts of this paper.

Funding Open access funding provided by National Institute for Health and Welfare (THL). This article grew from my PhD research at the University of Cambridge, funded by the Arts and Humanities Research Council; the British Society for the Philosophy of Science; Cambridge Commonwealth, European and International Trust and Newnham College.

References

Angner, E. (2011). Current trends in welafre measurement. In J. B. Davis & D. W. Hands (Eds.), *The Elgar Companion to Recent Economic Methodology*. Northampton: Edward Elgar.

Angner, E. (2013). Is it possible to measure happiness? *European Journal for Philosophy of Science, 3*(2), 221–240. https://doi.org/10.1007/s13194-013-0065-2

Baccelli, J. (2020). Beyond the metrological viewpoint. *Studies in History and Philosophy of Science Part A, 80*, 56–61. https://doi.org/10.1016/j.shpsa.2018.12.002

Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of Measurement. *Sociological Methods & Research, 9*(2), 147–160. https://doi.org/10.1177/004912418000900202

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

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

Brown, J. F. (1934). A Methodological Consideration of the Problem of Psychometrics. *Erkenntnis, 46*(1), 46–61.

Cartwright, N., Bradburn, N., & Fuller, J. (2016). *A Theory of Measurement* (No. 2016/07). Durham.

Casey, B. M., McIntire, D. D., & Leveno, K. J. (2001). The Continuing Value of the Apgar Score for the Assessment of Newborn Infants. *New England Journal of Medicine, 344*(7), 467–471. https://doi.org/10.1056/NEJM200102153440701

Chang, H. (2004). *Inventing Temperature: Measurement and Scientific Progress*. Oxford University Press.

Chang, H. (2009). Operationalism. In *Edward 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.

Decoene, S., Onghena, P., & Janssen, R. (1995). Representationalism under attack. *Journal of Mathematical Psychology, 39*(2), 234–242.

Embretson, S., & Reise, S. (2000). *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates Publishers.

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.

Fiske, D. W. (1971). *Measuring the concepts of personality*. . Aldine Pub. Co.

Galluzzo, G., & Loux, M. J. (2015). *Problem of Universals in Contemporary Philosophy*. . Cambridge University Press.

Heilmann, C. (2015). A New Interpretation of the Representational Theory of Measurement. *Philosophy of Science, 82*(5), 787–797. https://doi.org/10.1086/683280

Howell, D. C. (2010). *Statistical methods for psychology*. . Thomson Wadsworth.

Isaac, A. M. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy, 91*(4), 683–704.

Kline, P. (1998). *The new psychometrics: science, psychology and measurement*. . Routledge.

Krantz, D., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. I: Additive and polynomial representations*. Academic Press.
Le Noury, J., Nardo, J. M., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., & Abi-Jaoude, E. (2015). Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ (Clinical Research Ed.), 351*, h4320. https://doi.org/10.1136/BMJ.H4320
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. . Addison-Wesley.
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
Luce, R. D., Krantz, D., Suppes, P., & Tversky, A. (1990). *Foundations of Measurement, Vol. III: Representation, Axiomatization, and Invariance*. London and San Diego: Academic Press.
Mari, L. (2005). The problem of foundations of measurement. *Measurement, 38*(4), 259–266. https://doi.org/10.1016/J.MEASUREMENT.2005.09.006
Mari, L., Carbone, P., Giordani, A., & Petri, D. (2017). A structural interpretation of measurement and some related epistemological issues. *Studies in History and Philosophy of Science Part A, 65–66*, 46–56. https://doi.org/10.1016/J.SHPSA.2017.08.001
Marmodoro, A., & Yates, D. (2016). *The Metaphysics of Relations*. . Oxford University Press.
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., Torres Irribarra, D., & Wilson, M. (2016). On the philosophical foundations of psychological measurement. *Measurement, 79*, 311–320. https://doi.org/10.1016/j.measurement.2015.11.001
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
McLendon, H. J. (1955). Uses of Similarity of Structure in Contemporary Philosophy. *Mind, 64*, 79–95. https://doi.org/10.2307/2251045
Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*(3), 398–407. https://doi.org/10.1037/0033-2909.100.3.398
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. (2005). The logic of measurement: A realist overview. *Measurement, 38*(4), 285–294. https://doi.org/10.1016/j.measurement.2005.09.004
Michell, J. (2008). Is Psychometrics Pathological Science? *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 7–24.
Michell, J. (2020). Representational measurement theory: Is its number up? *Theory & Psychology*. https://doi.org/10.1177/0959354320930817
Narens, L., & Luce, R. D. (1993). Further comments on the "nonrevolution" arising from axiomatic measurement theory. *Psychological Science, 4*(2), 127–130.
Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. . McGraw-Hill.
Osherson, D., & Lane, D. (2018). Levels of Measurement. Retrieved August 5, 2018, from http://onlinestatbook.com/2/introduction/levels_of_measurement.html
Reiss, J. (2008). *Error in Economics: Towards a More Evidence-Based Methodology*. . Routledge.
Snaith, P. (1993). What Do Depression Rating Scales Measure? *British Journal of Psychiatry, 163*(3), 293–298. https://doi.org/10.1192/bjp.163.3.293
Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science (New York, N.Y.)*, *103*(2684), 677–680. https://doi.org/https://doi.org/10.1126/science.103.2684.677
Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* (pp. 1–49). Wiley.
Tal, E. (2012). *The Epistemology of Measurement: A Model-based Account*. University of Toronto.
Tal, E. (2016). Making Time: A Study in the Epistemology of Measurement. *The British Journal for the Philosophy of Science, 67*(1), 297–335. https://doi.org/10.1093/bjps/axu037
Velleman, P. F., & Wilkinson, L. (1993). Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. *The American Statistician, 47*(1), 65–72. https://doi.org/10.1080/00031305.1993.10475938
Vessonen, E. (2020). The Complementarity of Psychometrics and the Representational Theory of Measurement. *The British Journal for the Philosophy of Science, 71*(2), 415–442. https://doi.org/10.1093/bjps/axy032

Wolff, J. E. (2019). Representationalism in Measurement Theory. Structuralism or Perspectivalism? In *Understanding Perspectivism* (pp. 109–126). Routledge. https://doi.org/10.4324/9781315145198-7