Psychometrics versus Representational Theory of Measurement

Abstract: Erik Angner has argued that simultaneous endorsement of the representational theory of measurement (RTM) and psychometrics leads to inconsistency. His claim rests on an implicit assumption: RTM and psychometrics are full-fledged approaches to measurement. I argue that RTM and psychometrics are only partial approaches that deal with different aspects of measurement, and that therefore simultaneous endorsement of the two is not inconsistent. The argument has implications for the improvement of measurement practices.

1. Introduction

It is widely agreed that there are, broadly speaking, two approaches to measurement in the social sciences: representational theory of measurement (RTM) and psychometrics (Angner 2011, cf. Krantz 1991). Both are widely used, but their methodological connections are underexplored. On the one hand, there seems to be very little interaction and exchange between proponents of the two approaches (Judd & McClelland 1998; Krantz 1991). It also seems that researchers in different fields of social sciences operate with just one approach with little regard for the other (Angner 2011, 2013). On the other hand, it has been noted that the two approaches have potential to inform each other (Judd & McClelland 1998; Krantz 1991). Erik Angner (2008, 2009, 2011, 2013) is one of the few people who have explored the connections between RTM and psychometrics in recent years, which is why I will here focus on his contributions.

According to Angner (2011), RTM and psychometrics are incompatible alternatives. He writes that (2011, 124) “the simultaneous endorsement of the two approaches to measurement would lead to inconsistency.” This inconsistency claim can be understood in two ways. On the one hand, Angner argues that “since it is possible to satisfy the strictures imposed by the one approach to measurement without satisfying those imposed by the other, a measure that has been validated in accordance with the one approach has not necessarily been validated in accordance with the other.” The thought is that a simultaneous endorsement of the two approaches leads to a situation where a given measure both is and is not validated. And that is inconsistent. On the other hand, Angner (ibid. 131) argues that "[RTM] entails that an observable ordering satisfying certain axioms is necessary for measurement whereas the psychometric approach entails that it is not".[1] And it would be inconsistent to say that a given aspect both is and is not necessary for measurement.

Without knowing the details of the two approaches, we can already detect a silent assumption that underlies Angner's inconsistency claim: RTM and psychometrics are full-fledged approaches to measurement, that is, both approaches deal with conditions that are sufficient for measurement. If, on the other hand, the two approaches are partial in the sense that they deal with different non-sufficient, but necessary conditions of measurement, there is nothing inconsistent about endorsing both approaches simultaneously. In that case saying that a measure has been validated in terms of psychometrics but not in terms of RTM does not mean that the measure both is and is not validated. Rather, it means that one non-sufficient condition for full-fledged measurement has been addressed via psychometrics and that some other condition(s) has not been dealt with. Furthermore, if the psychometric approach deals with non-sufficient conditions of measurement, then the psychometric approach does *not* entail that the axiomatic conditions RTM deals with are not necessary for measurement. Rather, psychometrics is silent about other necessary aspects of measurement.

In this paper I argue that RTM and psychometrics do indeed focus on different aspects of measurement, both of which have to be dealt with in order for measurement to take place. Thus, instead of conceiving of RTM and psychometrics as full-fledged approaches to measurement, we should view them as partial approaches. If RTM and psychometrics solve different subproblems, we can establish what I call *the consistency claim*: simultaneous endorsement of RTM and psychometrics does not lead to inconsistency.

This argument does more than just disputes Angner's interpretation. If RTM and psychometrics are partial approaches to measurement, full-fledged testing of the measurement properties of a specific measurement instrument cannot be based solely on one of these approaches. In other words, if a measure has been scrutinized only in terms of one of these approaches, the results

[1] I will deal with Angner's contention that RTM requires an *observable* ordering in section 4.

from the usage of such an instrument do not count as measurements, unless further study of the instrument is conducted. And as we will see in the course of this paper, it is relatively common that certain social scientific measurement instruments are validated only in terms of the psychometric approach. The way I map out the scope of RTM and psychometrics helps us diagnose such problematic measurement practices and directs our choice of remedies. Towards the end of the paper I will suggest that many current validation practices might benefit from a complementary usage of RTM and psychometrics. Evidently a refutation of Angner's inconsistency claim is needed for such a proposal to be furthered.

I proceed as follows. Section 2 introduces RTM and psychometrics. Section 3 argues that RTM and psychometrics focus on different aspects, both of which are crucial for measurement. It then derives the consistency claim. Section 4 considers objections, and section 5 considers broader implications of the argument. Section 6 concludes.

2. Preliminaries

2.1 RTM

According to RTM, measurement involves "the construction of homomorphisms (scales) from empirical relational structures of interest to numerical structures that are useful." (Krantz et al 1971, 9, henceforth: FOM, from *Foundations of Measurement*). Homomorphisms are many-to-one mappings, and in RTM these mappings are from the empirical relational structures to numerical ones. To measure, one needs to prove two types of theorems. Firstly, one needs a representation theorem, which establishes that if a given empirical relational structure of interest satisfies certain (non-contradictory) axioms, then a homomorphism ϕ to a certain numerical structure can be established. Second, a uniqueness theorem establishes the permissible transformations of ϕ that also yield a homomorphism to the same numerical structure. Usually one distinguishes between four types of homomorphisms, i.e. scales: ratio, interval, ordinal and nominal. Ordinal scales, such as IQ, allow monotonic increasing transformations of the form $\phi \rightarrow f(\phi)$. Interval scales, e.g. temperature measured in Celsius or Fahrenheit, are such that they represent equality and inequality of intervals of the target attribute. For such scales, the permissible transformations are of the form $\phi \to \alpha\phi + b, \alpha > 0$. Ratio scales, such as length and weight, represent equality and inequality of intervals and have a non-arbitrary zero point. They allow for multiplicative transformation of the form $\phi \to \alpha\phi, \alpha > 0$.

In the RTM approach, measurement is based on empirical (relational) structures. In order to measure, we have to investigate empirically the relations between targeted objects, and establish that the empirical structure of interest satisfies the axioms that guarantee the existence of the mapping from the empirical structure to the numerical one. For example, the conditions that a empirical structure has to fulfill in order for it to be meaningfully represented on an ordinal scale are:

Let A be a 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 commodity bundles, and the relation $\succcurlyeq$ denotes a preference relation, i.e. $a \succcurlyeq b$ is interpreted as a is at least as preferred as b. If the testing of preferences reveals that the empirical relation $\succcurlyeq$ satisfies connectedness and transitivity, then one can prove a representation theorem: there is a function ϕ from A to the set of real numbers such that for all commodity bundles a and b in A, $a \succcurlyeq b$ iff $\phi(a) \geq \phi(b)$, that is, in informal terms, the preference 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.

2.2 Psychometric Validation

The problem with describing psychometric validation is that the concept has several meanings in contemporary methodological literature as well as in practice (Markus & Borsboom 2013). Here I shall operate with a focus on reliability and so-called construct validity, because that is how Angner describes psychometric validation (2011, 2013), and because it is a prominent way to go about psychometric validation.

On the psychometric approach, you start off by characterizing the target construct, i.e. the latent variable of interest, such as well-being or intelligence, and by proposing a measure (usually in the form of a questionnaire) of that construct. You then administrate the test and run a series of statistical tests on the response data to check whether the measure is reliable and has construct validity. Reliability amounts to the testing of the stability and consistency of the results the measure yields. There is a multitude of ways of doing this in practice, but it is common to check whether the test yields the same (or reasonably similar) result for a test taker when she takes it on another occasion (test-retest reliability) and to check whether the individual test items correlate with each other to a sufficient degree (internal consistency reliability) (Kline 1998, 29-30; Angner 2011, 128).[2]

[2] Angner (2011) uses Kline's (1998, esp. 78) specification of reliability, which explicitly assumes classical test theory rather than item response theory. (The difference between the two is, roughly, that classical test theory treats manifest scores as a function of true scores and error, while item response theory makes more detailed assumptions about the determinants of the manifest score.) Angner does not make this assumption explicit in his characterization of psychometric validation, which is why I do not discuss the assumption here extensively but rather accept it as an assumption that comes along with the specification of psychometric validation that Angner uses and that is, consequently, at issue here. Other specifications of the psychometric approach are discussed in sections 4 and 5 below.

Construct validation, on the other hand, is thought of as the test of the degree to which the measure captures the construct it is supposed to capture (Kline 1998; Angner 2011). In line with Cronbach and Meehl's (1955) seminal characterization of the process, the researcher should begin construct validation by formulating theoretical expectations of how the target construct relates to other constructs and measures, and then proceed to check whether the expected associations between these measures do indeed emerge. For example, suppose that our target construct, the unobservable *latent variable*, is well-being, and we have devised a questionnaire that purportedly captures this variable. If we have a theory that links well-being with mental health, then the construct validity of the new measure can be (partly) investigated by checking whether the purported well-being measure correlates to a sufficient degree with relevant measures of mental health.[3]

Consider, as an example, the famous Satisfaction with Life Scale (SWLS) that was devised to capture a particular target concept, i.e. unobservable *latent variable*: subjective well-being. SWLS consists of five questionnaire items, and asks subjects to rate their agreement with each of the items on a scale from 1 to 7. According to its authors Diener and colleagues (1985), the measure was validated, that is, it was shown to capture the target construct subjective well-being, when researchers compared responses on the SWLS to responses on other existing measures of subjective well-being and related constructs such as affect intensity, happiness and mental health. The results confirmed their expectation that SWLS scores correlate highly with those measures that also elicit a judgment on subjective well-being, but less so with measures that are intended to capture other related but distinct notions.

2.3 Radically Different Approaches

Angner (2011, 123) claims that RTM and psychometrics are "radically different". It is hard to disagree, given that RTM focuses on proving theorems while psychometrics focuses on

[3] It is a matter of debate as to what kind of theorizing can underwrite claims about construct validity. Angner hardly mentions theory in his characterization of psychometrics. Others, such as Alexandrova & Haybron (2016) propose that a more substantive theoretical framework is needed for psychometric validation to work.

relationships between different measurement instruments – two very different kinds of activities. One way to interpret this radical difference is to say that the two approaches deal with different, independently sufficient conditions for measurement. While Angner does not describe the two measurement approaches in terms necessary and sufficient conditions of measurement, it is clear from the way he writes that he treats them as independent and self-contained approaches, i.e. if you have one you do not need the other. For example, he argues (2011, 147) that under some circumstances in which RTM is inapplicable, psychometrics is "the only game in town" and therefore the only option for those who are keen on measurement. I take this claim to manifest the assumption that the two approaches are independent and self-contained. Such an interpretation is admittedly tempting in the context in which Angner makes his inconsistency claim. Angner (2008, 2009, 2011, 2013) argues that well-being researchers from economics and psychology have tended to rely on different approaches: orthodox welfare economists have often relied on RTM, whereas psychologists (and some heterodox economists) have validated their measures in terms of the psychometric approach. If that is indeed the case, it can be taken as evidence that in practice psychometrics and RTM are treated as two different full-fledged ways to go about measurement.

I believe there is a more fruitful way to think about the radical difference between RTM and psychometrics: they are partial approaches that deal with different non-sufficient but necessary aspects of measurement.[4] I call these aspects *representational interpretability* and *procedural validity*. These aspects are in many ways intertwined in practice, in particular, when procedural validation is done with extreme care the result is the fulfillment of the condition of representational interpretability. But the analytic distinction between the two aspects is nonetheless helpful for understanding what RTM and psychometrics can and cannot do. I argue that RTM focuses on representational interpretability but is silent about procedural validity, while the reverse is true for psychometrics.

[4] This does not mean that psychometrics or RTM are themselves necessary for measurement. There *may* be other ways of dealing with representational interpretability and procedural validity. I take it that my formulation of what constitutes representational interpretability and procedural validity are weak enough to accommodate a variety of ways for dealing with these aspects. Similar requirements for an adequate approach to measurement are expressed in Cartwright, Bradburn & Fuller (2016).

3. Partial Approaches

3.1 RTM and Representational Interpretability

Measurement is widely and almost invariably considered to involve numerical representation. The need for representational interpretability arises from the further observation that when it comes to measurement, not all numerical representations of empirical properties are created equal. The requirement of representational interpretability reflects the intuition that given certain empirical relations, some numerical representations are more appropriate than others to represent these relations, in the sense that they are appropriately interpretable in terms of the target system. But is representational interpretability a necessary aspect of measurement?

Consider a simple example. Lena has transitive strict preferences over slices of cakes: Black Forest $\succ$ Sacher $\succ$ Baked Alaska. If any numerical assignment would do, we could assign numbers to cakes as follows: Black Forest would be represented by -1, Sacher by 100 and Baked Alaska by 50. But assigning 3 to Black Forest, 2 to Sacher and 1 to Baked Alaska is informative about an interesting property of Lena's preferences, namely, order, which the former assignment fails to account for. If you agree that an adequate approach to measurement should be able to weed out the former assignment because it does not lend itself to a meaningful interpretation of the target system (preferences), you should agree that some kind of representational interpretability is crucial for measurement. That measurement requires representational interpretability may even seem obvious to you. But existing measurement practices show that its importance isn't always recognized: psychometricians are often accused of arbitrary and uninterpretable numerical representation of their target systems (FOM 33; section 3.3 below).[5] It is therefore important to consider what representational interpretability amounts to, and how RTM helps attain it.

What exactly does it take for a numerical representation to be interpretable in terms of the targeted empirical system? At the very least, a criterion of interpretability should weed out arbitrary numerical assignments. To achieve this, we could take our cue from S.S. Stevens'

[5] Psychometricians are not accused of uninterpretability that is as radical as in the example above, but the failures of interpretability they are accused of are, so to say, of the same type. See section 3.3 for more on this.

definition of measurement (e.g. 1975) and assign numbers to processes according to a rule. The trouble is that we need to specify the concept of a rule for it to get rid off arbitrary assignments. Stevens does not provide much help here, for he states that: "[t]he only rule not allowed would be random assignment…" (ibid. 47) But that just begs the crucial question, namely, how should we specify the notion of a rule so that random assignments are excluded. Stevens also speaks about the importance of matching operations as the basis of measurement, for example when people match numbers to sensations. But it is legitimate to compare the informativeness of numerical representations that different matching operations yield, suggesting that there is more to the interpretability of a numerical representation than just that it results from matching. (One could, after all, ask people to match numbers -1, 100 and 50 to objects so that -1 is assigned to the most preferred, 100 to second best option, and 50 to the least preferred one.)

There is another criterion for representational interpretability that is well-defined, intuitive and discriminates between alternative representations: a numerical structure has to *mirror (or map onto)* the empirical structure it is supposed to represent in order for it to be a useful representation.[6] This criterion immediately gets rid of the previous troublesome representation, because the suggested assignment of -1, 100 and 50 does not mirror the relevant empirical structure, namely, order. Here's another example: if we have four rods a, b, c and d such that when they are set side by side, the difference between the length of a and b is equal to that between c and d, a useful numerical representation mirrors this, so that $\phi(a)$ - $\phi(b) = \phi(c) - \phi(d)$. And so on for other relational structures. Representations that capture these mirroring relations are intuitive and useful exactly because they tell us how to interpret the numerical structure, and arithmetical operations on the assigned numbers, in terms of the target objects and their relations, when these objects and relations are examined in light of a given attribute. Thus

[6] The argument for the significance of "mirroring" (or "mappings", I use these terms interchangeably) is meant to be robust with respect to a variety of epistemological and metaphysical positions that are advocated in the measurement literature. In other words, the argument for thinking about mirroring as necessary for measurement meant to apply regardless of how one thinks about the nature and ontology of the objects that are being represented numerically and regardless of how one thinks we come to know about relationships between those objects. Consequently, I use "mirroring" as a thin or minimal notion, in the sense that filling in the details of what *exactly* a mirroring amounts to is left for the proponents of different epistemological and metaphysical accounts of measurement. (See also Tal 2015, esp. section 3.)

the notion of mirroring grounds representational interpretability better than the other candidates, i.e. rules and matching.

RTM builds on precisely this conception of representational interpretability, and studies meticulously the conditions under which such mirrorings or mappings can be said to hold. The whole point of the representation theorem is to establish the conditions under which a given relation between numbers that are used to represent the target system has a parallel relation in the realm of the objects, so that a given empirical relation exists between objects if and only if the numbers assigned to those objects have the corresponding relation to each other. The uniqueness theorem, in turn, establishes how the numbers in the numerical representation can be transformed without breaking the mapping between the empirical relations and the numerical ones. As indicated by the complexity of some of the axiomatic systems considered in FOM, as well as the serious intellectual effort that goes into proving the representation and uniqueness theorems, the conditions for the existence of such appropriately interpretable representations are not at all self-evident. RTM thus provides the conditions under which there is a rationale for a specific kind of numerical representation of an empirical structure and the transformations that preserve a representation of that structure. In doing so it deals with a crucial aspect of measurement: representational interpretability.

3.2 Psychometrics and Procedural Validity

The problem of validity of procedures stems from the observation that there is often a discrepancy between our best characterizations of the concept we want to measure and the reach of the empirical procedures that we want to use as tools for capturing that concept. This is because a target concept, that is, the aspect of an empirical system that we want to study, such as temperature, happiness or intelligence, has meaning independent of the procedures that are supposed to capture that concept. Although operationalism has been offered as a measurement strategy in the past, it is widely regarded as inadequate to simply define the target concept in terms of a procedure by stipulation. Recognizing that (many) concepts are not defined in terms of procedures gives rise to the problem of procedural validity: how do we know that a suggested procedure captures the concept we want it to capture, and how do we know it tracks empirical manifestations of this concept reliably across conditions. I take it to be evident that these are questions that a full-fledged approach to measurement should be able to address. It is hard to imagine measurement without adequate procedures.

In recent historico-philosophical literature on measurement, the proposed solution to the difficulty in validating procedures has been appeal to coherence (e.g. Chang 2004; van Fraassen 2008). The idea is, roughly, that claims about the appropriateness of a given procedure for the measurement of a given concept requires multiple determinations of the same concept via different fallible methods, or multiple determinations of related concepts via different methods. If the different determinations agree with each other that tends to argue in favor of the assumption that the procedures are indeed capturing the target concepts. In other words, the hypothesis that a measure is capturing what it is intended to capture gains evidence from the fact that multiple determinations of that construct (or theoretically related constructs) cohere with each other.

It is easy to see that the psychometric approach, in particular the strategy of construct validation, is an instance of a coherentist approach to the validation of a measurement procedure (Alexandrova & Haybron 2016). Construct validation starts from a web of theoretical assumptions that entail that a proposed measure of the target concept correlates with certain other measures (and across contexts). The hypothesis is tested by checking whether such correlations emerge, and if they do, that is taken as evidence for the claim that the proposed measure indeed captures the correct construct. In other words, the claim that the proposed measurement instrument captures the correct construct is supported by appeal to coherence between relevant measures and theoretical expectations. Under the above characterization of coherentism in measurement, construct validation counts as a coherentist solution to the problem of validity of procedures.

3.3 Limits

We have seen that RTM deals with representational interpretability and psychometrics deals with procedural validity. Let me now discuss how each of these approaches is silent about the task that the other one focuses on. Start with RTM and procedural validity. Many scholars think that RTM is unhelpful when it comes to finding an appropriate measurement procedure (Boumans 2004; Reiss 2008). Julian Reiss (2008, 67) puts the point as follows: "[RTM] tells us what kind of structure an attribute or a phenomenon must have in order to be measurable *given we have a reliable measurement instrument*, but it does not tell us where and how to look for a reliable instrument in the first place." Reading through the authorative statement of RTM supports this observation: FOM says virtually nothing about procedures. Even when the authors consider empirical examples of applications of the axiomatic conditions that FOM explores (e.g. transitivity), they do not identify how the applicability of a certain axiomatization in a given empirical context can be established.[7] As the authors of FOM themselves note, their empirical examples have the modest role of "motivating" and "illuminating" the axiomatic foundations the book is primarily concerned with (FOM, xvii).

To get a clearer grasp of this, consider the paradigmatic example of RTM: measurement of length with rigid rods. Even this case does not tell enough about procedures to guarantee that the axioms apply. This is because simple observations of differences between lengths of rigid rods cannot be considered reliable, when these differences are extremely small. We need more subtle procedures to state that the axiomatic conditions hold in extreme circumstances. Similarly for another seemingly simple case of RTM, namely, weight. Suppose we place two objects on an equal-arm pan balance and the arms remain horizontal. How do we judge that the two objects are indeed equal in weight, rather than that our balance is broken or simply not sensitive enough to detect the difference between the two? Although the first chapter of FOM mentions some of these issues, the axiomatizations that the rest of the three volumes deal with do not solve, or purport to solve them. Note that none of this is to say that tracking the axiomatic conditions that

[7] This should *not* be taken as a criticism of the authors of FOM. Their focus on axiomatizations rather than procedures does not imply that they *failed* to acknowledge the necessity of procedures for measurement. Rather they chose to focus on axiomatizations rather than procedural considerations, because their aim in FOM was not to give a full-fledged account of measurement but to lay down the representational *foundations* of measurement. All this is compatible with my argument, for the aim is not to criticize RTM but to show its scope.

RTM lays down does not *require* procedures but rather that RTM only gives advice on what the relevant conditions are, not how they can be captured by means of measurement procedures. RTM can hardly be an approach to validating procedures if it is virtually silent about procedures.

How does the psychometric approach deal with representational interpretability? Several authors have noted that in psychometrics, the appropriateness of a given type of numerical representation (usually interval level) is often assumed rather than established (Borsboom & Zand Scholte 2008; see also Kristoffersen 2010; Hobart et al 2007). This points to neglect for the question of representational interpretability. Such neglect is implicit in practices where subjects are asked to rate their standing on some attribute (e.g. well-being) on a pre-specified scale, and these scores[8] are taken to constitute an interval scale of the target attribute, implying (by definition of interval scale) that differences between the assigned numbers mirror equality and inequality of distances between objects on the measured attribute. But it is not trivial that differences between the assigned numbers represent distances between objects on the measured attribute. Subjects may or may not be using the rating scale so that this assumption is fulfilled, and usually they aren't (Hobart et al 2007). By taking the scores as interval level measurements, psychometricians assume that the emerging numerical representation mirrors certain relations between the objects of study when they are compared in terms of the target attribute. Thus they operate under the assumption that the numbers have a mirroring relation to manifestations of the underlying attribute, but neglect rigorous study of the conditions for the existence of such mirroring relations, and empirical testing of whether or not those conditions are fulfilled by the target system.

The assumption concerning interval level measurement is left unsubstantiated, because the validation techniques described in our specification of the psychometric approach (section 2.2) are not apt for ensuring evidence of representational interpretability. Test-retest reliability and internal consistency reliability give us evidence of how test results correlate across time and how individual items correlate with each other, but they don't tell us anything about how we should

[8] Similar considerations apply to normalized ratings (or normed scores). For more on how the above-discussed issues manifest vis-à-vis normalized ratings, see Blanton & Jaccard 2006, esp. 36-37.

interpret equalities and inequalities in manifest scores in terms of the latent attribute. Similarly, while construct validation tells us that the proposed measurement procedure in some sense captures the correct construct (rather than a related one), it does not tell us whether the emerging numerical representation is interpretable in the sense that the appropriate mirroring relation exists, at least when it comes to interval and ratio scales.[9] Knowing that two measures are related in some way and to some extent does not (and is not meant to) ensure that either measure yields interval level measurement. In fact there is a substantial psychometric literature (starting with Stevens 1951) that argues that interval level measurement is a *precondition* for a meaningful interpretation of many of the statistical tests that are commonplace in the construct validation exercise. (About the permissibility of statistical tests, see Stevens 1951; Luce 1959; cf. Hobart et al 2007; Kristoffersen 2010). The psychometric approach therefore builds on assumptions about representational interpretability but does not help establish the truth of those assumptions.

3.4 Consistency, finally

We have seen that RTM gives guidance on representational interpretability but is silent about procedures, while the psychometric method of construct validation gives a coherentist response to procedural validity but is silent about conditions for representational interpretability. I have also argued that both representational interpretability and procedural validity are crucial aspects of measurement. Thus we can conclude that RTM and construct validation are only partial approaches to measurement.

This reveals that there is in principle no reason to believe that "the simultaneous endorsement of the two approaches to measurement would lead to inconsistency" as Angner (2011) claims. To say that a measure has been validated in terms of psychometrics but not in terms of RTM does not mean that the measure both is and is not validated. It means that one condition for full-fledged measurement has been satisfied via psychometrics and that another aspect has either not been studied or has been appropriately dealt with relying on some other approach than RTM.

[9] Arguably construct validation does lend support to the claim that the data is ordinal, cf. Ferrer-i-Carbonell & Frijters 2004, 643. But usually psychometricians want to operate with interval scales.

Furthermore, the fact that the psychometric approach deals with non-sufficient conditions of measurement means that the psychometric approach does *not* entail that the axiomatic conditions RTM outlines are not necessary for measurement. Rather, the psychometric approach is just silent about an aspect of measurement that RTM deals with. It is therefore consistent to simultaneously endorse psychometrics and RTM.

4. Objections

Let me now discuss potential objections. First of all, it could be argued that Angner intends his inconsistency claim merely as a characterization of how practicing well-being researchers (and social scientists more broadly) treat (or have treated) these two approaches, not as a general claim about their incompatibility. In that case the argument of this paper should be framed as a critique of failed practices, not Angner. But I think there is ample evidence in Angner (2011) that he at least sometimes means the latter, more general and stronger claim, and is therefore the proper target of this counterargument. Firstly, Angner (2011, section 6.2) describes the two measurement approaches *qua* methodologies instead of *qua* practices, and makes the inconsistency claim on the basis of these methodological characterizations, not on the basis of observations of practice. Second, it seems that a claim about the inconsistency of RTM and psychometrics needs to incorporate some general, non-practice-based considerations, because as Angner (2011, 147) himself acknowledges, many social scientists (in particular economists) are not aware of the existence of two measurement approaches and therefore do not conceptualize their measurement activities in terms of these approaches. The upshot is not that we cannot make claims about how RTM and psychometrics manifest in practice but that claims about the logical compatibility of the two have to build on some generalized characterizations. This is because arguably claims about logical inconsistency require some explicit characterizations of the things that are claimed to be inconsistent, and in this case practice does not supply those characterizations. So it seems to me that Angner's inconsistency claim and my consistency claim need to be interpreted as general claims about RTM and psychometrics, albeit ones that rely partly on evidence from social scientific practice and that have implications for measurement practices in social sciences.

What about objections to my characterizations of the two measurement approaches? It may be objected that I have misrepresented psychometric validation and that in fact psychometric validation does include considerations of interpretability. In response I reiterate that psychometric validation has multiple definitions in the vast psychometric literature (see Markus and Borsboom 2013). Representing psychometric approach in terms of reliability and construct validation reflects Angner's account, and I think Angner's characterization captures much of psychometric practice, although not all of it. Some psychometricians do worry about (what I have called) representational interpretability. The way they usually study this aspect is by testing the extent to which their data fit so-called Item Response Theory (IRT) models (more on these in section 5 below). Crucially for the present purposes, when psychometricians do discuss representational interpretability, they clearly distinguish methods for dealing with this aspect (e.g. goodness-of-fit tests with IRT models) from construct validation (and reliability as described above) (Hobart et al 2007; Blanton & Jaccard 2006). This strengthens my claim that the psychometric approach, as described here, is not an approach to establishing representational interpretability. Furthermore, these authors argue that representational interpretability is often neglected in psychometrics, which enforces the point that most of psychometric measurement ignores representational interpretability.

This brings me to another objection, namely, that construct validation cannot possibly be called an appropriate solution to procedural validity if it cannot guarantee that a measurement procedure yields an interpretable numerical representation. It is true that ultimately we do want procedural validation to yield an appropriately interpretable numerical representation. But sometimes, as with construct validation, validation of a procedure only tells us that a measure captures the correct construct, not how exactly the resulting numbers reflect the target attribute. More specifically, construct validation gives evidence for the appropriateness of weak claims about representational interpretability (e.g. claims about ordinal data), but fails to establish representational interpretability in the maximally informative way that psychometricians require in order to make claims about interval level measurement. Blanton & Jaccard (2006) observe this when they argue that psychometric measures are often valid but at the same time arbitrary, in the sense that it is not known how a one-unit change in the observed scores reflects changes on the underlying dimension. Strictly speaking, then, it is more accurate to call construct validation a partial solution to validation of procedures, because it only establishes some aspects of the adequacy of a procedure.

Turn to RTM now. I have implicitly assumed that the distinction between observable and unobservable is not relevant for RTM. But Angner (2011, 2013) claims that RTM applies only to observable orderings and structures. Reinterpreting RTM in this way might look like I am changing the subject rather than genuinely challenging Angner's claim. But my argument holds whether or not RTM is taken to apply to unobservables. If RTM applies only to observables, it is partial in two ways: it does not tell about procedural validity, and it only deals with the representational interpretability of observable target attributes. Because RTM and psychometrics are still only partial approaches, the consistency claim is left unaffected.

Why diverge from Angner's interpretation then? Firstly, and most importantly, I believe the authors of FOM did not endorse Angner's strict interpretation. Krantz et al write that "[t]he axioms purport to describe relations, perhaps idealized in some fashion, among certain *potential* observations" (FOM, 26-27, italics added). Sometimes observations do not conform to the axiomatic conditions because of the inability of the experimental setting to adequately capture the target phenomenon. One possible solution according to FOM is to consider relational statements such as $\square \succcurlyeq \square$, not as statements about observations, but as theoretical statements inferred from the data (Suppes et al 1989, 300). This suggests that observability of the fulfillment of the axiomatic conditions is not strictly required. (In any case, the dividing line between observables and unobservables is notoriously contested.) Second, Heilmann (2015) has recently argued that the theorems of RTM can be readily and usefully applied to relations that have no empirical (let alone observable) content. The restrictive view of RTM would exclude such useful applications, and potentially others as well, such as complementary usage of RTM and psychometrics.

5. Implications

I have argued that simultaneous endorsement of RTM and psychometrics is consistent, because RTM and psychometrics are not full-fledged approaches to measurement. The direct consequence of this argument is that Angner is wrong to claim that simultaneous endorsement of RTM and psychometrics leads to inconsistency.

To avoid the impression that all of this was said just to criticize Erik Angner, let me consider what the broader implications of my argument are vis-à-vis measurement in social sciences.[10] Thus far psychometricians have largely ignored the kind of abstract mathematical measurement theory that RTM embodies (Cliff 1992), whereas proponents of RTM have been openly suspicious of the extent to which psychometrics counts as measurement (FOM 33). There has thus been little interaction between the two approaches. But in light of the recognition that procedural validity and representational interpretability are necessary and intertwined aspects of measurement, the potential benefits of increased interaction between proponents of the different approaches become apparent. Given the specializations of proponents of RTM and psychometricians, the intersection of considerations over procedural validation and research on representational interpretability is likely to be fruitful grounds for a dialogue between these two approaches.

The above call for exchange in measurement expertise may sound like a fluffy "let's all be friends"-conclusion. But the stakes are actually high in medical and social scientific measurement. Psychometric measures of welfare, health, educational achievement and other social scientific constructs are frequently used to inform policy-making and decision processes that have significant impact on people's lives. For example, UK Treasury has started to explore the possibility of using psychometric measures of subjective well-being to help determine which public policies to fund (Fujiwara & Campbell 2011), and clinical trials of antidepressants have for decades employed psychometric measures to establish the effectiveness of drugs (Bagby & al

[10] I believe (but do not have the space to argue here) that the argument about the necessity of representational interpretability and procedural validity extends beyond social sciences. However, it could be that their necessity is acknowledged better in physical sciences, and thus the problem of incomplete validation does not emerge there in the same way it emerges in social sciences. Naturally, the claim about the shortcomings of psychometric validation has no consequences for measurement in physical sciences, because psychometrics is a social scientific approach.

2004). It is hardly acceptable that such decisions are made on the basis of measures that lack representational interpretability, but in fact the appropriateness of both subjective well-being measures and depression rating scales has been questioned on these grounds (Kristoffersen 2010; Bagby & al 2004). These are forceful reasons for psychometricians to explore RTMs area of expertise. It is likely that similar considerations go for fields that approach measurement from the perspective of RTM, and these would constitute additional reasons for exploring complementary usage.

How would joint usage of RTM and psychometrics actually look like? We have already touched upon an area of research that seems to manifest the potential for complementary usage of RTM and psychometrics, namely, research on IRT models, in particular, the Rasch model.[11] It has been argued that the Rasch model, which psychometricians sometimes use to establish whether or not the data is interval data, is a probabilistic instantiation of one of the axiomatic structures that RTM promotes, namely, additive conjoint measurement (see Borsboom & Mellenbergh 2004). In other words if the data fit the Rasch model reasonably well, that is thought to show that we have interval level measurement because a fit to the Rasch model indicates that certain axiomatic conditions for representability are fulfilled. The jury is still out on whether or not (and how) the Rasch model instantiates additive conjoint measurement, but at the very least these ongoing research efforts illustrate the potential for points of convergence for psychometrics and RTM. While this is not the place to explore these points of convergence further, they are worth mentioning here due to their connection to the consistency claim I have advanced. Acknowledging the limited scope of RTM and psychometrics and the consistency of their simultaneous endorsement are crucial first steps to fruitful research in these likely areas of complementary usage of RTM and psychometrics.

[11] The Rasch model is the simplest of IRT models. It specifies a relationship between the probability of a given response to an item, characteristics of test items (e.g. the difficulty of an individual item) and the latent attribute of interest, such as aptitude. The model can be expressed as follows: $P_i(\theta) = \frac{\exp(\theta - \beta_i)}{1+\exp(\theta - \beta_i)}$, where $P_i(\theta)$ is the probability of a correct response to item i from a randomly selected examinee whose ability level is θ, and β_i is the item difficulty parameter. See e.g. Embretson & Reise 2009.

6. Concluding remarks

I have argued that simultaneous endorsement of RTM and psychometrics is consistent, because RTM and psychometrics are not full-fledged approaches to measurement. The immediate implication of this is that doubt should be cast upon claims that are advanced under the assumption that RTM and psychometrics are full-fledged approaches and the assumption that simultaneous endorsement of RTM and psychometrics is inconsistent. If practicing social scientists use either approach as if it is full-fledged and self-contained, as seems to be the case in many fields, those practices need to be scrutinized in terms of the partial nature of the adopted measurement approach and remedied so that all of the necessary aspects of measurement are taken care of. More positively, the claim about the partial nature of RTM and psychometrics points to ways in which the two approaches can inform and even complement each other.

Bibliography

Alexandrova, A. and D. Haybron. 2016. "Is Construct Validation Valid?" *Philosophy of Science* 83(5).

Angner, E. 2013. “Is it Possible to Measure Happiness? The Argument from Measurability.” *European Journal for Philosophy of Science* 3(2): 221–240.

2011. “Current Trends in Welfare Measurement.” In John B. Davis and D. Wade Hands (Eds.) *The Elgar Companion to Recent Economic Methodology*. 121–154. Northampton: Edward Elgar.

2009. “Subjective Measures of Well-Being: Philosophical Perspectives.” In H. Kincaid & D. Ross (eds.) *The Oxford Handbook of Philosophy of Economics.* Oxford: Oxford University Press.

2008. “The Philosophical Foundations of Subjective Measures of Well-Being.” In L. Bruni, F. Comim & M. Pugno (eds.) *Capabilities and Happiness*. Oxford: Oxford University Press.

Bagby RM, Ryder AG, Schuller DR, Marshall MB. 2004. "The Hamilton Depression Rating Scale: has the gold standard become a lead weight?" *American Journal of Psychiatry 161*(12): 2163–77.

Blanton, H. and J. Jaccard. 2006. “Arbitrary metrics in psychology.” *American Psychologist 61*(1): 27-41.

Borsboom, D., & G. J. Mellenbergh. 2004. “Why Psychometrics is Not Pathological A Comment on Michell.” *Theory & Psychology*, *14*(1):105-120.

Borsboom, D. and A. Zand Scholten. 2008. “The Rasch model and conjoint measurement theory from the perspective of psychometrics.” *Theory & Psychology 18*(1): 111-117.

Boumans, M. 2005. *How Economists Model the World into Numbers*. New York: Routledge.

Cartwright, N., Bradburn. N and Fuller. J. 2016. “A Theory of Measurement.” Durham University: CHESS Working Paper No. 2016-07

Chang, H. 2004. *Inventing Temperature: Measurement and Scientific Progress.* Oxford: Oxford Univ. Press.

Cliff, N. 1992. “Abstract Measurement Theory and the Revolution That Never Happened.” *Psychological Science* 3(3):186–90.

Cronbach, L. J., & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological bulletin*, *52*:4, 281.

Diener, E., R. Emmons, R. Larsen and S. Griffin. 1985. “The satisfaction with life scale.” *Journal of personality assessment 49*(1): 71-75.

Embretson, S.E. and Reise, S.P., 2009. *Item response theory for Psychologists*. New York: Psychology Press.

Ferrer‐ i‐ Carbonell, A., & Frijters, P. 2004. “How Important is Methodology for the estimates of the determinants of Happiness?*.” *The Economic Journal*, *114*(497):641-659.

Fujiwara, D. and R. Campbell. 2011. *Valuation techniques for social cost-benefit analysis: stated preference, revealed preference and subjective well-being approaches.* HM Treasury Green Book Discussion Paper.

Heilmann, C. 2015. “A New Interpretation of the Representational Theory of Measurement.” *Philosophy of Science* 82(5): 787-797.

Hobart, J., S. Cano, J. Zajicek and A. Thompson. 2007. “Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations.” *The Lancet Neurology* 6(12): 1094-1105.

Judd, C. and G. McClelland. 1998. “Measurement.” In Fiske, Susan, Daniel Gilbert and Gardner Lindzey (eds.). *Handbook of social psychology, 4th edition*. Boston: McGraw-Hill.

Kline, P. 1998. *The New Psychometrics: Science, Psychology, and Measurement.* London: Routledge.

Krantz, D. 1991. From Indices to Mappings: The Representational Approach to measurement. In Brown, D. & Smith, J. E. (eds.) *Frontiers of Mathematical psychology. Essays in Honor of Clyde Coombs.* New York: Springer.

Krantz, D., R. D. Luce, A. Tversky, and P. Suppes. 1971. *Foundations of Measurement Volume I: Additive and Polynomial Representations.* San Diego and London: Academic Press.

Kristoffersen, I. 2010. “The metrics of subjective wellbeing: Cardinality, neutrality and additivity.” *Economic Record*, *86*(272): 98-123.

Luce, R. D. 1959. "On the Possible Psychophysical Laws." *Psychological Review,* 66, 81-95.

Markus, K. and D. Borsboom. 2013. *Frontiers of test validity theory: Measurement, causation, and meaning*. New York: Routledge.

Reiss, J. 2008. *Error in Economics. Towards a more evidence-based methodology*. New York: Routledge.

Stevens, S. 1975. *Psychophysics: Introduction to Its Perceptual, Neural and Social Prospects*. New York: Wiley & Sons.

1951. "Mathematics, Measurement, and Psychophysics." In S. S. Stevens (ed.) *Handbook of Experimenital Psychology.* New York: John Wiley.

Suppes, P., D. Krantz, R. D. Luce, and A.Tversky. 1989. *Foundations of Measurement Vol 2: Geometrical, Threshold and Probabilistic Representations*. San Diego and London: Academic Press.

Tal, E. 2015. "Measurement in Science", The Stanford Encyclopedia of Philosophy (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/measurement-science/>.

van Fraassen, Bas. 2008. *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.