Accepted for publication in the British Journal for the Philosophy of Science

Preprint

The Complementarity of Psychometrics and the Representational Theory of Measurement

Elina Vessonen

Abstract

Psychometrics and the representational theory of measurement (RTM) are widely used in social scientific measurement. They are currently pursued largely in isolation from one another. I argue that despite their separation in practice, RTM and psychometrics are complementary approaches, because they can contribute in complementary ways to the establishment of what I argue is a crucial measurement property, namely, Representational Interpretability. Because RTM and psychometrics are complementary in the establishment of Representational Interpretability, the current separation of measurement approaches is unfounded.

1 Introduction

Measures of social scientific concepts such as aptitude, well-being, and depression play a key role in determining life-changing decisions, from university admissions to drug approval and national economic policies. These measures had better be good to warrant their status as life-changers. But the scientific experts are divided on the question: how do we determine whether a social scientific measure is valid or not? Different views on what determines measure validity imply divergent commitments on deep philosophical questions, such as questions pertaining to observation, representation, modeling and theorizing. For instance, conceptions of measure validity often specify conditions under which an instrument's readings provide an adequate representation of the measurand. [1] They may also specify what kinds of relations between theory, models and observations allow inferences concerning measure validity. [2] The study of approaches to social scientific measurement therefore links directly with issues that have traditionally concerned philosophers of science.

There are, broadly speaking, two approaches to measurement in social sciences: psychometrics and the representational theory of measurement (RTM).[3] Roughly, the psychometric approach focuses on testing associations between tests that usually have the form of a questionnaire, while RTM conceives of measurement in terms of so-called representation and uniqueness theorems. Both approaches are widely used, but they are pursued largely in isolation of one another. Proponents of RTM infrequently venture into employing psychometric techniques, and psychometricians are rarely aware of the existence of RTM.

Such a separation is hardly surprising when viewed in light of the history of RTM. The authors who lay the ground for RTM, as expressed in the three-volume *Foundations of Measurement*[4], were partly motivated by their disappointment with psychometrics, which had

[1] For example (Suppes and Zinnes [1963], section 2.3).
[2] For example (McClimans *et al* [2017]; Alexandrova and Haybron [2016]).
[3] See for example (Angner [2011]; Judd and McClelland [1998]).
[4] (Krantz *et al* [1971]; Suppes *et al* [1989]; Luce *et al* [1990]).

dominated social scientific measurement at least since 1930s.[5] In the beginning of the first volume of *Foundations of Measurement,* philosopher Patrick Suppes and psychologists David Krantz, R. Duncan Luce and Amos Tversky express discontent with psychometrics, arguing that it is 'far from clear' how to interpret the measurement results psychometric instruments yield. Elsewhere Suppes complains that psychometrics consists of an 'array of bewildering and conflicting catechisms'.[6] According to Suppes it is unclear how following these catechisms is supposed to result in genuine measurement, whereas RTM provides a non-dogmatic framework for social scientific measurement.

But psychometricians have been averse to RTM, and little of Suppes' and other's work has been adopted by psychometricians. According to Norman Cliff ([1992]), the main reason is that the abstract mathematical apparatus that RTM employs is foreign to psychometricians. Cliff argues that it has been difficult for psychometricians to see how the general mathematical principles of RTM can be made to apply to specific empirical issues. Some philosophers of science enforce this conclusion, claiming that RTM is too narrow to be useful for the practical execution of measurement.[7] Nonetheless RTM seems to have resonated well with economists, who use formal techniques abundantly. In a sense RTM and economics were allied from the start, because Suppes and others frequently mention economists' attempts to measure utility as examples of RTM.[8]

The fact that two approaches have not meshed in practice does not imply that they are incompatible in principle. Alas, there is no consensus on the interrelations between RTM and psychometrics. The literature that explores such connections sends mixed messages about the compatibility of the two. For example, philosopher Erik Angner ([2011]) has argued that the two are incompatible in such a way that a simultaneous endorsement of them would lead to inconsistency. By contrast, in their introduction to social scientific measurement, Charles Judd and Gary McClelland [1998] indicate that RTM and psychometrics could in principle inform and even complement each other. But complementarity is not the focus of their article, and the proposal has not been pursued further.

In this paper I deliver on the vague suspicion of complementarity. I argue that despite their isolation in practice, RTM and psychometrics are complementary. By this Complementarity

[5] This conservative estimate is based on the fact that the Psychometric Society was founded in 1935.
[6] (Suppes and Zinnes [1963], p. 3).
[7] (Mari *et al* [2017]; Boumans [2016]; Reiss [2008], chapter 4).
[8] For example (Krantz *et al* [1971], p. 9; Luce and Tukey [1964]).

Claim I mean that RTM and psychometrics can contribute in different but interlocking ways to the establishment of a measurement property I call Representational Interpretability. Roughly speaking, Representational Interpretability is the requirement that relations between numbers are interpretable in terms of empirical relations in the measured system. I argue that RTM contributes to Representational Interpretability by establishing conditional statements of the form: if conditions *x, y,* and *z* are fulfilled, then we have an interpretable numerical representation. Techniques within psychometrics contribute to Representational Interpretability by providing evidence for or against the antecedent of the conditional. The upshot of the argument is that RTM and psychometrics are complementary. While this does not mean that psychometrics and RTM have to be *always* used in conjunction with one another (for example, measurement in the physical sciences proceeds fine without psychometrics), the Complementarity Claim should urge us to rethink the current separation of the two approaches.

Measurement has been increasingly on philosophers' radar. With literatures on general measurement theory and measurement in the physical sciences already well established,[9] philosophers have grown increasingly interested in the challenges of social scientific measurement (see for example McClimans et al [2017]; Isaac [2017]; Alexandrova [2017]; Alexandrova and Haybron [2016]; Cartwright et al [2016]; Boumans [2016]; Heilmann [2015]; Hood [2015]; Angner [2011]; Reiss [2008]). The present paper contributes to philosophical measurement scholarship by explicating and arguing for the importance of Representational Interpretability and by laying out the roles RTM and psychometrics can play in its fulfillment. In particular, the argument engages with philosophers' recent claims about the practical limitations of RTM,[10] showing what RTM – when appropriately interpreted – has to offer for social scientific measure validation.

The paper proceeds as follows. Section 2 characterizes RTM and psychometrics. Section 3 explains the Complementarity Claim conceptually. Section 4 details an example of complementarity. Section 5 concludes with potential objections and practical implications of the argument.

[9] See for example (Chang [2004]; van Fraassen [2008]; Frigerio *et al* [2010]; Soler *et al* [2013]; Riordan [2015]). Tal ([2013]) provides a historical overview.
[10] (Mari *et al* [2017]; Boumans [2015]; Reiss [2008, ch. 4]).

2 Two Approaches to Measurement

2.1 RTM

The RTM approach received its canonical statement in *Foundations of Measurement*, which was written by philosopher Patrick Suppes and psychologists R. Duncan Luce, Amos Tversky and David Krantz. 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], p. 9). Homomorphisms are many-to-one mappings. In RTM these mappings are from the empirical relational structures to numerical ones. To measure, one needs to prove two types of theorems. A representation theorem 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. A uniqueness theorem establishes the permissible transformations of φ that also yield a homomorphism to the same numerical structure.

It is common to distinguish at least three types of homomorphisms, i.e. scales: ratio, interval, and ordinal.[11] Ordinal scales, such as Mohs hardness scale for minerals, allow monotonic increasing transformations of the form $\phi \rightarrow f(\phi)$. Such transformations preserve order relations, as the name of the scale suggests. Interval scales, e.g. temperature measured in Celsius, represent equality and inequality of intervals of the target attribute. For such scales, the permissible transformations are of the form $\phi \rightarrow \alpha\phi + b, \alpha > 0$. Ratio scales, such as length, represent equality and inequality of intervals and have a non-arbitrary zero point so that equalities and inequalities of ratios are meaningful. Ratio scales allow for multiplicative transformation of the form $\phi \rightarrow \alpha\phi, \alpha > 0$. The latter two scale types are usually called quantitative or cardinal.

In the RTM approach, measurement is based on empirical relational structures, which are, roughly, sets of relations that can hold between entities in the target domain. In order to measure, one has to establish the fulfillment of relational constraints that guarantee the existence of a mapping from an empirical structure to a numerical one. These constraints are stated in axioms. For example, the axioms pertaining to ordinal scales are:

[11] On other scale types, see (Suppes and Zinnes [1963]).

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 commodity bundles, and the relation $\succcurlyeq$ denotes a preference relation, i.e. $a \succcurlyeq b$ means a is at least as preferred as b. If 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 $\phi(a) \geq \phi(b)$. 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, ϕ' is a permissible transformation of ϕ as long as it preserves the order of the numbers assigned to the objects.

Interpretations of RTM are debated in philosophical literature (Tal [2012]). For example, there is disagreement on whether RTM applies to unobservable attributes (Vessonen [2017]; Heilmann [2015]; Angner [2011]; Mari [2000]) and whether RTM is an epistemological or a merely formal theory of measurement (Tal [2012]). I will come back to some of these debates, but for now I will use this summary: a) according to RTM, measurement requires the establishment of homomorphisms between empirical and numerical structures, and b) RTM delivers on this requirement in the form of representation and uniqueness theorems. This summary stems from the canonical *Foundations of Measurement*, where the focus is almost exclusively on proving theorems to establish homomorphisms. Some proponents of RTM would agree that measurement requires more than just proving theorems. That in no way contradicts my summary of RTM. The point is that RTM itself focuses on and delivers representation and uniqueness theorems.[12]

[12] My interpretation of RTM bears similarities to those expressed in (Heilmann [2015]; Narens and Luce [1993]).

2.2 Psychometrics

The concept of psychometric validation has several meanings in contemporary literature (Markus and Borsboom [2013] provide an overview). The following characterization captures the big picture, ignoring subtleties. On the psychometric approach, one starts off by characterizing the target construct, i.e. the latent variable of interest, such as well-being or aptitude, and by proposing a measure, usually in the form of a questionnaire, of that construct. One also offers a specification of independent variables that are taken to be relevant for explaining or predicting the observed test scores (dependent variable), and specifies an assumption or a model of how those independent variables combine to predict the observed score. For example, the Classical Test Theory (CTT) approach assumes that the observed test score is a function of so-called true test score and measurement error. CTT is a dominant approach to thinking about the determinants of the observed score (Engelhard [2013]), but several alternative models are gaining currency within psychometrics.

Once these assumptions about the target construct have been made, one proceeds to administer the test and run a series of statistical tests on the response data to check whether the measure has the desirable properties that a validated measure is supposed to have. To give an example, some tests are meant to determine the extent to which assumptions about the determinants of the observed score hold. Another category of tests involves checking for so-called construct validity, which is thought of as the test of the degree to which the measure captures the construct it is supposed to capture. To establish construct validity, one checks that the results of the proposed measure converge with results from other measures that are theorized to capture the same target attribute and that the results diverge from measures that have other attributes as their theorized targets (see Cronbach and Meehl [1955]; Campbell and Fiske [1956]). There is a multitude of other tests of validity, the details of which do not concern us here.

2.3 Representational interpretability

RTM and psychometrics conceive of measurement in different ways: RTM focuses on proving theorems while psychometrics focuses on statistical tests on empirical data. It is not surprising that such seemingly different perspectives on measurement have not meshed. As mentioned, the developers of RTM frequently framed RTM as the correct way to do social scientific measurement, which they thought was unjustifiably dominated by psychometricians' 'bewildering and conflicting catechisms' (Suppes and Zinnes [1963]; see also Krantz et al [1971], ch. 1). But RTM did not resonate

with psychometricians, partly because it seemed so far removed from empirical applications (Cliff [1992]; Luce [1997]; see also Mari et al [2017]; Reiss [2008]). Thus psychometrics carried on largely unaffected by mathematical theories of measurement, and RTM found its allies in economics, where proving representation and uniqueness theorems is part and parcel of measurement practice. The disciplinary divide has prompted some philosophers to claim that the two approaches are in some sense incompatible (Angner [2011]).

Despite this seeming incompatibility of the two, and the historical divide between their proponents, I argue that RTM and psychometrics are methodological complements. By this I mean that they contribute different types of evidence to the establishment of a crucial measurement property. The insights each supplies complements the one provided by the other.

To make this argument I will focus on what I call the requirement of Representational Interpretability, which I define as follows:

> Representational Interpretability$_1$ (RI1): The requirement for Representational Interpretability is fulfilled if and only if specified relationships between numbers assigned to entities have an interpretation in terms of relationships between those entities, when those entities are compared in terms of the target attribute.

The definition is a mouthful, but the idea is simple. Consider a situation where we have evidence that Maya has transitive strict preferences over cakes: Lemon $\succ$ Strawberry $\succ$ Chocolate. Here the relevant entities are cakes. They are being compared in terms of preferences, which is the target attribute. When thus compared, the relevant empirical relation between the cakes is ordering. In order to be Representationally Interpretable in this case, a numerical representation has to capture the (preference) ordering of the cakes by assigning numbers to cakes such that the numbers have the correct ordering relation to each other, for example Lemon $\rightarrow$ 3, Strawberry $\rightarrow$ 2 and Chocolate $\rightarrow$ 1.

As the example suggests, the intuition behind Representational Interpretability is simple: representation of preference order requires ordered numbers. Its fulfillment in practice, however, is complicated. Firstly, we want to represent more complex empirical structures than ordering. The complexity of the structures increases as we expand the class of represented relationships, which in turn increases the difficulty of establishing Representational Interpretability. Second, Representational Interpretability is an empirical requirement. We need evidence that the

relevant relations in fact exist in the target system – we cannot just stipulate this. For example, for the ordered numbers to be interpretable in terms of Maya's preferences, evidence from, say, stated preference questionnaires should reliably show that Maya indeed has the transitive preferences expressed in the example. But if different questionnaires yield contradicting evidence, the fulfillment of Representational Interpretability is not clear. Such underdetermination need not worry us now though, for the aim is not to provide an all-purpose manual for handling contradicting evidence, but to zoom in on the roles RTM and psychometrics can play in establishing Representational Interpretability.

That measurement requires something like Representational Interpretability is acknowledged implicitly and explicitly across measurement scholarship.[13] When a psychometrician assumes that the target attribute is measured on a certain scale something like the requirement of Representational Interpretability is usually evoked implicitly. Furthermore, it is implicit in the set-up of RTM that Representational Interpretability is a crucial component of measurement, but the founders of RTM did not explicate the concept. Its significance might have also been ignored because the formalism of RTM is off-putting and impenetrable to many. Either way, Representational Interpretability has not been previously explicated in an accessible way. Consequently, the complementary roles of RTM and psychometrics in the establishment of this property have not been appreciated.

The requirement of Representational Interpretability fits neatly with a recent philosophical theory of measurement proposed by Cartwright *et al* ([2016]). They include representation as a requirement, focusing on representation theorems and the coordination of how an instrument indicates manifestations of the target attribute. I endorse these requirements, but want to add that there is more to representation than proving representation theorems and coordinating the indicator–attribute relationship. The additional step requires us to collect reliable evidence showing that manifestations of the target attribute in the target domain instantiate the conditions for representation on the scale type of interest.[14] That step establishes what I call Representational Interpretability.

[13] See (Michell [1993]) on the history of representational ideas of measurement.

[14] In practice the activities for establishing Representational Interpretability interconnect with coordination of the attribute-indicator relationship. For present purposes it suffices to zoom in on the former.

How do RTM and psychometrics contribute to Representational Interpretability? I argue that RTM provides a formal characterization of the conditions for a meaningful interpretation. In other words, RTM shows that if conditions *x, y* and *z* are fulfilled, then there is a numerical representation that is interpretable in terms of the empirical system that fulfills those conditions. On the other hand, some psychometric techniques supply empirical and statistical evidence for (or against) the antecedent of the conditional statement. The Complementarity Claim (CC) can be summarized as follows:

> Complementarity Claim (CC): RTM and psychometrics are complementary approaches in the sense that they contribute in different, interlinking ways to the body of evidence that establishes a crucial measurement property called Representational Interpretability.

Before we embark on the argument that leads to the Complementarity Claim, let me emphasize that while I believe that Representational Interpretability is necessary for measurement, it is likely not sufficient. To be sufficient, the conditions of measurement would have to include, for example, a more detailed account of the aforementioned coordination of the indicator-attribute relationship. This account might specify how the target attribute (causally) interacts with the measurement instrument so as to bring about an interpretable numerical representation. Since the aim here is not a comprehensive theory of measurement (and since these matters have been discussed extensively elsewhere, for example Chang [2004]; Tal [2016]; Cartwright et al [2016]; McClimans et al [2017]), but rather an account of how RTM and psychometrics can complement each other, we can safely zoom in on Representational Interpretability, while simultaneously acknowledging that there is more to measurement than that.

The argument leading to the Complementarity Claim is complicated, which is why I divide the discussion into three parts. First, I explain the Complementarity Claim conceptually. Second, I show how complementarity looks like in practice. Third, I return to philosophical literature on social scientific measurement and embed my argument in it.

3 Complementarity, Conceptually

3.1 RTM: Conditions of representational interpretability

Measurement is 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 are created equal. The requirement of Representational Interpretability reflects the intuition that some numerical representations are more appropriate than others to represent certain empirical relations.

Why is Representational Interpretability necessary for measurement? Take our simple example from before, Maya and her transitive strict preferences over cakes: Lemon $\succ$ Strawberry $\succ$ Chocolate. If any numerical assignment would do, we could assign numbers to cakes as follows: Lemon $\rightarrow -1$, Strawberry $\rightarrow 100$ and Chocolate $\rightarrow 50$. But assigning 3 to Lemon, 2 to Strawberry and 1 to Chocolate is informative about an interesting property of Maya's preferences, namely, order, which the former assignment fails to account for. If you agree that measurement should be able to weed out the former assignment because it doesn't lend itself to a meaningful interpretation of Maya's preferences, you should agree that some kind of Representational Interpretability is crucial for measurement.

What exactly does it take for a numerical representation to be interpretable in terms of the targeted empirical system? A plausible suggestion is that a numerical structure has to mirror the empirical structure it is supposed to represent. In the example of Maya's preferences, mirroring means that the assigned numbers have such an ordering relation to each other that it reflects the preference ordering of the cakes. But there are other, more complex structures. For 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 the difference between the numbers assigned to *a* and *b* is equal to that between numbers assigned to *c* and *d*. For example, the assignment $a \rightarrow 4$, $b \rightarrow 3$, $c \rightarrow 6$, $d \rightarrow 5$ works but $a \rightarrow 4$, $b \rightarrow 3$, $c \rightarrow 6$, $d \rightarrow 2$ doesn't.

As many philosophical theories of scientific representation, these examples of mirroring capitalize on our intuitions of structural similarity: ordering of numerals mirrors ordering of entities, equalities of numbers mirrors equalities of entities (in terms of an attribute) and so on. But a general definition of representation in terms of similarity is notoriously elusive.[15] Fortunately, common scale types allow us to enumerate some similarity relations that ground interpretations of numbers in terms of the targeted empirical structure without recourse to a general definition of

[15] See for example (van Fraassen [2008]; Isaac [2013]).

representation. The most common mirrorings that allow intuitive interpretations of numbers in terms of entities, and that thus imply Representational Interpretability, are ones where: i) order relations between numbers map onto order relations between entities (ordinal scales), ii) (in)equalities of differences between numbers map onto (in)equalities of differences between entities (interval scales), and iii) (in)equalities of ratios between numbers map onto (in)equalities of ratios between entities (ratio scales) when entities are compared in terms of the degree to which they manifest the target attribute. There are other scale classifications and thus other interesting mirrorings, but the most common ones suffice presently.

The representation and uniqueness theorems of RTM are crucial for establishing the conditions under which such mirrorings hold. The whole point of the representation theorem is to show what conditions an empirical relational system has to fulfill in order for it to map onto a numerical system of interest. The uniqueness theorem, in turn, establishes how the numbers in the numerical assignment can be transformed without breaking the mapping between the empirical relations and the numerical ones, thus establishing the relevant scale type. In supplying the axiomatic conditions for representation on the above-enumerated familiar scales, RTM sets forth constraints for mirrorings between empirical and numerical systems, and thus contributes to the establishment of Representational Interpretability.

Let us summarize. The axiomatizations RTM supplies are conditions for certain mirrorings between empirical and numerical systems. Mirrorings, in turn, are our grounds for interpreting numerical assignments in terms of empirical systems, that is, they ground Representational Interpretability. Moreover, knowledge of mirrorings is crucial for knowing what kinds of arithmetic (and statistical) operations can be meaningfully applied to the numbers. For example, taking the arithmetic mean of ordinal values is not meaningful because arithmetic mean is not defined for ordinal scales. The same applies to more complicated operations (see Stevens [1951]; Luce [1959]). For present purposes, we need not settle what scales allow for what operations. Rather the take home is this: established mirroring relations ground Representational Interpretability, allowing us to weed out uninformative numerical representations (such as my first suggestion for presenting Maya's cake preferences) and meaningless operations (such as taking the arithmetic mean of ordinal values).

3.2 Psychometrics: Evidence of representational interpretability

At this stage we should note that RTM does not tell us anything about how to establish that particular axiomatic conditions are fulfilled. Consider the axiom of transitivity, which is a necessary conditions for an ordinal representation: If $a \succcurlyeq b$ and $b \succcurlyeq c$, then $a \succcurlyeq c$. The axiom does not (and is not meant to) tell us how to establish that the transitive relations hold in a given empirical situation. In fact, the axiom itself doesn't even tell us how to interpret the relationship $a \succcurlyeq b$. a and b might be rods or commodities or people, and the operator $\succcurlyeq$ might be interpreted as "at least as long as" or "at least as preferred as" or "at least as well off as". Even when the axioms are given in an interpreted form, the question of establishing that such conditions hold in a particular empirical situation is something that representation and uniqueness theorems cannot solve. It is one thing to argue that preferences can be represented on an ordinal scale if and only if preferences are transitive, and quite another to argue that stated preference questionnaires show reliably that people's preferences in fact are transitive.

These considerations are meant to crystallize that there are two sides of Representational Interpretability: one is establishing conditional statements concerning the conditions under which an interpretable numerical representation holds, and another one is establishing that those conditions hold in a specific context. My Complementarity Claim rests on the observation that some (but not all!) psychometric models instantiate axiomatic conditions expressed in RTM. That is, although a psychometric model is never stated in terms of axioms, some models postulate structures that embed assumptions about conditions of Representational Interpretability. Tests of fit between data and such models can therefore act as evidence that we have an interpretable numerical representation of the target system. In summary:

> P1: If structural empirical conditions expressed in axioms x, y and z are fulfilled then we have an ordinal/interval/ratio level representation of the attributes of interest.[16]
>
> P2: If the attributes of interest have the structure postulated in psychometric model P, then manifestations of the attribute fulfill structural empirical conditions x, y, and z.
>
> P3: If empirical tests of fit between data and psychometric model P show that the data fits the model, then we have evidence that the attributes of interest have the structure postulated in psychometric model P.[17]

[16] Sometimes this premise is a biconditional statement.

[17] A more precise formulation of the argument in terms of evidence is given in section 4.2.

Conclusion$_1$: If the data fit model P, we have evidence that the attributes of interest are representable on an ordinal/interval/ratio scale.

RI2: The requirement for Representational Interpretability is fulfilled if and only if the attributes of interest are representable on the scale type of interest.[18]

Conclusion$_2$. If the data fit model P, we have evidential support for the fulfillment of the requirement of Representational Interpretability.

The conclusion of the argument is not the Complementarity Claim. But the argument clinches the Complementarity Claim in this sense: when we fill in the argument outline with supporting evidence for each premise, we can see that the axiomatizations of RTM and certain psychometric models have interlocking roles to play in supporting Representational Interpretability. The representation and uniqueness theorems of RTM constitute proofs for statements that have the form expressed in P1. P2, when suitably filled in, is true by virtue of a demonstrable mathematical connection between a particular psychometric model and a particular axiomatization. P3 is an empirical claim about the potency of statistical tests of how well the psychometric model fits the data. RI2 expresses more compactly the idea captured in RI1. [19]

To put flesh to the bones of this argument, we will study a psychometric model known as the Rasch model. There is a small technical literature within psychometrics that explores whether or not, and in what sense the Rasch model is an instantiation of an RTM-style axiomatic structure known as conjoint measurement (Michell [2014]; Kyngdon [2011]; Embretson and Reise [2000]; Borsboom and Mellenbergh [2004]). I will now explicate the nature of the connection between Rasch and conjoint measurement and use it to drive home the Complementarity Claim.

4 Complementarity in Action

4.1 What is the Rasch model?

[18] The scale type of interest depends on the researcher and the questions she aims to answer.

[19] Since scales are distinguished informally according to the kinds of relations a numerical representation is informative of, we can shorten RI1 to RI2. For more on the formal definition of scales in terms of uniqueness properties, see Suppes and Zinnes ([1963]).

The Rasch model is one of several models known as Item Response Theory (IRT) models.[20] It was first proposed by Danish mathematician Georg Rasch ([1960]). Rasch, like other IRT models, is often thought of as an alternative to Classical Test Theory (CTT), which has been and is the standard of psychometric testing at least since 1930s (Embretson and Reise [2000], 13). The Rasch model, like all IRT models, explains the probability of a correct response to a test question (or the probability of a specific response, when dealing with multiple response categories per item)[21] as a result of the influence of the ability level[22] of the examinee and characteristics of the test items, where 'characteristic' denotes things like the difficulty of the item.

Rasch is the simplest IRT model: it includes one item characteristic, namely, difficulty, while other models include parameters for characteristics such as item discrimination (how informative the item is of different ability levels) and susceptibility to guessing (how likely low-ability examinees are to guess the correct response). The more complex models are often needed, because typically test performance cannot be explained (predicted) only in terms of ability and the difficulty of the item. If the response data does not "fit" the simplest model, more complex models may be examined.

There are two widely used versions of the Rasch model, one that treats probability of the correct response as a dependent variable and one that treats log-odds of a correct response as the dependent variable. The first Rasch model specifies the following relation:

$$P_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)}$$

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.

[20] I follow Embretson and Reise ([2000]) and Engelhard ([2013]), although some authors specify IRT so that Rasch is not part of these models.

[21] I discuss dichotomous items, i.e. ones that are either correct/incorrect or endorse/not endorse.

[22] The notion of 'ability' is used here for convenience, but Rasch does apply to attributes that are perhaps less obviously 'abilities'.

When the model is used in psychometric practice, the data that is collected using the psychometric instrument of interest is tested against the Rasch model. The first step is estimating item and ability parameters from the data, that is, estimating the difficulty of the individual items and the abilities of examinees. A popular method for doing this is joint maximum likelihood method. The details of this process do not matter much for present purposes, for our focus is on goodness-of-fit tests between the Rasch model and data.[23] Roughly speaking, the joint maximum likelihood method estimates the ability of an examinee by investigating the likelihood of her response pattern (correct/incorrect on each item) conditional upon different levels of latent ability. The same kind of estimation is done for the item difficulty parameter, and the re-estimation of the two parameters is done iteratively until neither parameter changes in two consecutive estimation steps.[24]

Once the parameters have been estimated, one tests the fit between the Rasch model and the data. To that end, a battery of statistical tests is imposed on the estimates that joint maximum likelihood method yields. For example, one goodness-of-fit test divides the examinees to ability groups based on the ability estimates, and compares the observed response patterns of each group with the predictions of the Rasch model (see Hambleton et al [1991], 59; Embretson and Reise [2000], ch. 9). In simplistic terms: one plugs in a given ability value *a* and a given item difficulty level *d* into the Rasch model and then proceeds to check whether the thus computed probability of correct response matches the actual frequency with which individuals of ability *a* gave the correct answer to an item that is of difficulty *d*. If the data fits the model well in the sense that the predictions of the model and the observed response patterns converge to a degree that is judged reasonably high, that is taken as evidence that the attribute of interest has the attribute structure postulated in the Rasch model. Importantly, lack of fit between the model and the data means that either the test is not appropriate or the hypothesized model of the target attribute is not appropriate. Either way, further study needs to be conducted to achieve measure validity.

While the focus of this paper is the Rasch model, one must keep in mind that validation of a psychometric measure does not end with Rasch analysis. By scrutinizing some measurement properties, Rasch contributes pieces to the larger validity puzzle, while remaining silent on other aspects. Tests of construct validity, for example, are also an important part of a full-fledged

[23] For more on parameter estimation, see (Embretson and Reise [2000]; Hambleton *et al* [1991]).
[24] The epistemic benefits of iteration are famous in philosophy of measurement. See (Chang [2004]).

validation exercise.[25] These other methods need not concern us here though, for our focus is a single desirable property – Representational Interpretability – and Rasch's role in establishing it.[26]

In the next section, I argue that tests of fit with the Rasch model provide evidence of Representational Interpretability, because they test the fulfillment of requirements for measurement on an interval scale. These requirements are set forth in the so-called theory of conjoint measurement, which is a brainchild of the RTM approach and which I will introduce in the next section. Because the Rasch model is an instance of the psychometric approach while conjoint measurement is an instance of the RTM approach, the case of the Rasch model demonstrates the complementarity of RTM and psychometrics.

4.2 Rasch and conjoint measurement

The theory of conjoint measurement (also known as additive conjoint measurement or simultaneous conjoint measurement) is one of the most celebrated axiomatizations in the RTM tradition. It was first proposed by psychologist R. Duncan Luce and statistician John Tukey in 1964. They were motivated by the fact that axiomatizations of cardinal measurement structures required that the attribute of interest allows side-by-side combination (so-called concatenation), but psychological attributes do not allow for such operations. We can compare the length of rigid rods by observing them side-by-side but we cannot set Andy's happiness next to Bobby's happiness and compare them. Luce and Tukey's proposed an axiomatization of so-called conjoint measurement, which achieves interval level measurement of attributes that do not allow for concatenation.

The axioms of conjoint measurement describe an empirical structure in which one attribute can be described as the simultaneous 'effect'[27] of two other attributes. More precisely,

[25] There is even a Complementarity Claim II to be made here, because it can be argued that construct validation techniques contribute to the establishment of a valuable measurement property that neither RTM nor the Rasch model can establish. While RTM and the Rasch model help determine measurement level, construct validation techniques help determine that the correct denotation of the target construct is captured i.e. not that the target entities are represented at the appropriate measurement level but rather that what is being represented (at whatever level) is relations between entities in terms of the correct target attribute. The explication of this other valuable measurement property, and the potential complementarity between RTM, the Rasch model and psychometric construct validation techniques, deserve a paper of their own.

[26] The Rasch model can fulfil other functions, which are not relevant for the present argument.

[27] I follow Luce and Tukey ([1964]) in using this term.

the axioms describe a situation where an attribute Y, which is the joint effect of component factors D and A, is the sum of the effect of A as captured by real-valued function ϕ and of the effect of B as captured by real-valued function ψ, i.e. Y = ϕ(D) + ψ(A) – in other words the component factors combine additively to form the joint effect. Luce and Tukey [1964] use the example that loudness of a tone can be thought as the effect of frequency and intensity of the tone.

The Rasch model is another example of the kind of attribute structure conjoint measurement corresponds to, because it posits that the probability (or log-odds) of a correct response to a test item can be thought of as the effect of difficulty of the question and the ability of the respondent. The fact that Rasch model instantiates the additive attribute structure can be readily seen from the log-odds version of the Rasch model:

$$\ln \frac{(P_{is})}{(1 - P_{is})} = \theta_s - \beta_i$$

where

P_{is} is the probability of a correct response to item i from subject s,

θ_s is the ability level of subject s, and

β_i is the item difficulty parameter (see Embretson and Reise [2000], 148).

This mathematical connection between Rasch and conjoint measurement was first noted by Keats ([1967]) and since then several people have discussed it (for example Perline et al [1979]; Andrich [1988]; Wright and Stone [1999]; Borsboom and Mellenbergh [2004]; Kyngdon [2008]; Embertson and Reise [2000]; Michell [2014]; Bond and Fox [2001]). However, the subject is not well-known and exists in the margins of psychometric literature. There's also ambiguity about the sense in which Rasch 'instantiates' conjoint measurement.[28] It is hard to find a thorough explication of the relationship in the literature, which is why I shall supply one here.

[28] Some people have disputed that the Rasch model instantiates conjoint measurement (for example Kyngdon [2008]). I believe the disagreement stems from different readings of "instantiates", which is why a thorough explication is needed here.

For our purposes, the most important axioms of conjoint measurement are the so-called cancellation axioms, and I shall focus on them.[29] This narrower focus is warranted, first, because the literature treats these axioms as the crucial targets of empirical testing of the conjoint structure (Embretson & Reise [2000], 148–149; Luce et al [1991], ch. 21.8). Second, to show the complementarity of RTM and psychometrics, we need both approaches to contribute to the establishment of Representational Interpretability, and empirical study of the fulfillment of some axioms is a contribution to that aim, even if we couldn't scrutinize all the axioms via the Rasch model. Third, it is possible that the Rasch model contributes evidence concerning the fulfillment of Solvability and the Archimedean axiom, but these contributions cannot be discussed independent of the details of the specific attributes under scrutiny (e.g. is the test measuring e.g. mathematical ability, quality of life etc.). This is because these axioms specify assumptions about the kinds of values each attribute can take. It is widely agreed that these axioms do not allow direct empirical testing but are rather accepted or rejected on the basis of 'general considerations' (see Luce et al [1991], section 21.8.4). Finally, the axiom of Weak order, involving conditions such as transitivity, is so weak that it is easy to see that the additive structure postulated in Rasch instantiates it.

To get a grasp of the cancellation axioms, consider Table 1, where each column corresponds to a value of one of the component factors (e.g. letters a, b, c are values for the component factor 'ability'), rows correspond to values of the other component factor (here letters d, e, f can be thought of as values for the component factor 'item difficulty') and the ordered pairs in each cell correspond to the joint effect (e.g. probability of correct response). The so-called single cancellation axiom requires that the relative order of levels of the effect attribute for any two levels of one component factor is the same regardless of the level of the second component factor. Table 1 illustrates this axiom: the order of the cells for two values of the column variable (a and b) is the same for all levels of the row variable (d, e, f), as indicated by the sameness of the direction of the arrows:

[29] The presentation of the axioms differs slightly from article to article. I present the axioms in a form I take to be most common in contemporary literature (see e.g. Kyngdon [2008]; Embretson and Reise [2000]).

	a	b	c
d	(d, a) →	(d, b)	(d, c)
e	(e, a) →	(e, b)	(e, c)
f	(f, a) →	(f, b)	(f, c)

Table 1. a, b, c are levels of one of the component attributes while d, e and f are levels of the other component attribute. The ordered pairs in the cells represent levels of the effect attribute.[30]

Another important axiom, known as double cancellation axiom, is often expressed graphically as in Table 2. The verbal interpretation is that if the order relations indicated by the dashed arrows hold, then the order relation indicated by the solid arrow must also hold.

	a	b	c
d	(d, a)	(d, b)	(d, c)
e	(e, a)	(e, b)	(e, c)
f	(f, a)	(f, b)	(f, c)

Table 2. Interpretation of columns, rows and cells as in Table 1.

From the axiomatization that Luce and Tukey give, they arrive at their seminal conclusion concerning interval level representability:

> From the axioms we give, simultaneous measurement on interval scales is obtained for each kind of quantity separately and for their joint effects. (Luce and Tukey [1964], 2):

In other words their seminal representation and uniqueness theorems showed that if the axioms (only two of which have been presented here) are fulfilled, the component attributes as well as the effect attribute have a meaningful interval level representation. (For the full axiomatization and the theorems see Luce and Tukey [1964]).

[30] To keep tables readable, not all arrows have been drawn.

With these axioms and representational results at hand, we get a more detailed account of the connection between conjoint measurement and the Rasch model. Beyond noting the sameness of the attribute structure, a further illustration of the connection between Rasch and conjoint measurement can be given by plugging in any item and ability levels in the Rasch model and calculating probability values to complete a two-way table, as in Table 3 and Table 4. It is easy to see that the above-described cancellation axioms are fulfilled when the Rasch model fits perfectly.

		Ability			
		−1	**0**	**1**	**1,5**
Item Difficulty	**−1**	0.50 →	0.73	0.88	0.92
	0	0.27 →	0.50	0.73	0.82
	0,5	0.18 →	0.38	0.62	0.73
	1	0.12 →	0.27	0.50	0.62

Table 3. Demonstration of the single cancellation axiom with the Rasch model. The cells present probabilities calculated from the Rasch model. Interpretation of arrows as in Table 1.

		Ability			
		−1	**0**	**1**	**1,5**
Item Difficulty	**−1**	0.50	0.73	0.88	0.92
	0	0.27	0.50 ↗	0.73 ↗	0.82
	0,5	0.18 ↗	0.38 ↗	0.62 ↗	0.73
	1	0.12 ↗	0.27 ↗	0.50	0.62

Table 4. Demonstration of the double cancellation axiom with the Rasch model. Interpretation of arrows as in Table 2.

We have now seen what is meant by the claim that the Rasch model instantiates the attribute structure of conjoint measurement: when the attributes have the structure postulated in the Rasch model, levels of the three attributes form the patterns postulated in the experimentally (dis)confirmable axioms. On grounds of Luce and Tukey's theorems, we also know that conjoint measurement yields an interval level representation. Following this, goodness-of-fit tests with the Rasch model can act as tests of whether a specific target attribute has the structure that allows it to be represented on an interval scale. How? Recall that the goodness-of-fit tests check whether the data conforms to the predictions of the Rasch model. As we see from Tables 3 and 4, the Rasch model predicts that the data forms the kinds of patterns that fulfill the cancellation axioms of conjoint measurement. So the better the predictions of the Rasch model converge with patterns in the data, the more we have evidence that the target attribute fulfills the cancellation axioms. This in turn is evidence for interval level representability. In summary:

> $P1_{RASCH}$: If we have evidence that the axioms of conjoint measurement are fulfilled then we have evidence of an interval level representation of the attributes of interest.
>
> $P2_{RASCH}$: If we have evidence that the attribute of interest has the structure postulated in the Rasch model, then we have evidence that manifestations of the attributes fulfill the axioms of conjoint measurement.
>
> $P3_{RASCH}$: If empirical tests of fit between data and the Rasch model show that the data fits the model, then we have evidence that the attributes of interest have the structure postulated in the Rasch model. [31]
>
> RI2: The requirement for Representational Interpretability is fulfilled if and only if the attributes of interest are representable on the scale type of interest.
>
> $Conclusion_{RASCH}$: If the data fit the Rasch model, we have evidential support for the fulfillment of the requirement of Representational Interpretability.

Note that the truth of $P3_{RASCH}$ is trivial conceptually speaking, because the whole point of goodness-of-fit tests is to inform us whether or not the attribute has the relevant structure. Empirically

[31] Assuming that the measure has been validated in other respects, e.g. that the data pertains to the attribute of interest (e.g. mathematical ability) rather than a different one (e.g. reading comprehension). Construct validation is the standard way of addressing this aspect.

speaking, though, a good fit to the Rasch model does not clinch the case for the attributes having the postulated structure. There are a variety of reasons for this, for example, the procedures of model parameter estimation and the goodness-of-fit tests have their own shortcomings and implementation-related difficulties (Embretson and Reise [2000]; Hambleton *et al* [1991]). Furthermore, testing the fit between data and the Rasch model means that a number of postulations about the target attribute are being tested simultaneously, and it is notoriously difficult to disentangle the source of the failure to fit the model (Borsboom and Mellenberg [2004]). That is why we need other methods to strengthen our conclusions concerning Representational Interpretability. For example, Kyngdon ([2011]) proposes tests that are meant to overcome blind spots of standard goodness-of-fit tests.

All this is to hammer the hopefully obvious point that Rasch, like any validation technique, is not all-encompassing. That does not undermine the present argument. The point is not that Rasch is fool-proof method for testing axioms, but that Rasch constitutes a psychometric technique for generating evidence for (or against) the empirical instantiations of structures postulated in the RTM approach. In doing so, it shows the complementarity of psychometrics and RTM in the establishment of Representational Interpretability.

5 Conclusion: Critics and Fruits of Complementarity

I have argued that Representational Interpretability is an important measurement property and that RTM and psychometric techniques can complement each other in establishing this property. I argued for the complementarity of RTM and psychometrics by showing that both approaches play a crucial role in a valid argument leading to a conclusion about Representational Interpretability. To show what complementarity looks like, I argued that usage of the psychometric model known as the Rasch model yields evidence of Representational Interpretability, because it acts as a test of conditions of conjoint measurement, which is an instance of the RTM approach.

One might be tempted to object that there is an inconsistency in my description of the measurement literature: on the one hand I have claimed that the interrelations of RTM and psychometrics are largely unexplored, on the other hand I have said that the mathematical connection between the Rasch model and conjoint measurement has been known at least since 1967. Doesn't the latter claim undermine the former?

No. Both claims are well-founded, because in practice the establishment of the connection between the Rasch model and conjoint measurement has not translated into a recognition of the complementarity of RTM and psychometrics. Why? Firstly, the connection between the Rasch model and conjoint measurement is still poorly known outside the theoretical literature on the Rasch model (see Perline et al [1979]). We can see this from the fact that when psychometricians use the Rasch model, they rarely justify its usage with reference to the connection between the Rasch model and conjoint measurement. For example, in a widely cited paper on the validity of the HAM–D measure of depression, the authors Bagby *et al* ([2004]) use evidence from the Rasch model to argue that the measure is not valid, but they say nothing about how the connection between Rasch and conjoint measurement is relevant for interpreting the results of the test. This is unsurprising: of course a practicing psychometrician cannot stop to scrutinize the theoretical underpinnings of each statistical method she uses. While poor understanding of the connection between Rasch and conjoint measurement need not undermine inferences about the validity of an instrument, it is an obvious obstacle to the acknowledgement of the complementarity of RTM and psychometrics.

Second, the Rasch model is by no means mainstream psychometrics, because the Classical Test Theory approach still dominates much of psychometric practice (see McClimans *et al* [2017]; Engelhard [2013]). The relatively limited usage of the Rasch model makes it understandable that the mathematical connection between Rasch and conjoint measurement has not trickled down into a general recognition of the complementarity of psychometrics and RTM.

Moving on to other objections. Some contributors argue that, contrary to my treatment of RTM, RTM is not just a formal framework laying out axiomatic conditions of numerical representation. Rather, RTM also sets restrictions on the kind of evidence that can vouch for the fulfillment of the axioms (Michell [1986]). In particular, some philosophers argue that only direct observations of the fulfillment of the axiomatic conditions are permissible evidence according to RTM (Michell [1986]; Mari [2000]; Angner [2011]). One way to characterize the difference between the presently used interpretation and the alternative, observability-tied interpretation is this: on the former interpretation, RTM is a 'theory of measurement' laying out representation-related formal structures, while on the latter RTM is a 'strategy of measurement' that sets out conditions for something to count as successful measurement, where some of those conditions pertain to permissible empirical evidence for representation.

The alternative take on RTM gives rise to two objections to the Complementarity Claim. First, one might object that the Complementarity Claim fails *tout court* because it rests on an incorrect interpretation of RTM. If we adopt the observability-tied interpretation of RTM, the argument goes, then RTM and psychometrics are *not* complementary, because most of the attributes psychometricians are interested in cannot be directly observed in the way RTM (allegedly) requires. Second, the Complementarity Claim can be said to fail as a counterproposal to claims about the incompatibility of RTM and psychometrics (for example the claims made in Angner [2011]), because to first change the operative interpretation of RTM and then argue against the incompatibility claims amounts to a change of subject (or even a fallacy of equivocation) rather than a genuine challenge.

Start with the first objection. Although distilling the *correct* interpretation of RTM from the literature is tricky,[32] there seems to be ample textual evidence that the authors of *Foundations of Measurement* did not believe that RTM requires direct observational evidence. Krantz and others write that '[t]he axioms purport to describe relations, perhaps idealized in some fashion, among certain *potential* observations' (Krantz *et al* [1971], 26–27, italics added). Sometimes observations do not conform to the axiomatic conditions because of the shortcoming of the experimental setting. One possible solution according to the authors is to consider relational statements such as $a \succcurlyeq b$, not as statements about observations, but as theoretical statements inferred from the data (Suppes et al [1989], 300). Elsewhere Suppes states that in the RTM approach, scales are defined entirely independent of the procedures that inform us about the empirical relational system of interest (Suppes and Zinnes [1963]; 15): 'Precisely what empirical observations are involved in the empirical system is of no consequence.' The upshot is that the first objection – claiming that the Complementarity Claim rests on an incorrect interpretation of RTM – does not hold water.

The second RTM-related objection avoids the slippery debate about correctness and instead argues that it is fallacious to challenge claims about the incompatibility of RTM and psychometrics by simply changing what one takes RTM to stand for. In response, note firstly that switching to another interpretation of RTM is a legitimate way to challenge incompatibility claims, if one can show that the presently proposed interpretation is better – that is, more accurate and/or more expedient – than the one that underwrites the incompatibility claim. I have already argued

[32] Several people have contributed to laying the foundations of the approach, and there's no guarantee that the exact specification of the approach is constant across individuals or across an individual's career.

that the present interpretation of RTM resonates better with the writings of the founders of RTM. In addition, other authors have argued that a broad interpretation of the applicability of the theorems of RTM can reap epistemic benefits, which the narrow observability-tied view cannot harvest (Heilmann [2015]). Furthermore, proponents of the alternative take on RTM need to draw a line between observable and unobservable relational structures in order to police illegitimate applications of RTM. The place of that dividing line is notoriously contested (see for example van Fraassen [1980]; Hacking [1983]). By contrast, a proponent of the present interpretation of RTM need not get entangled in that thorny debate. Hence there are both correctness- and expediency-related reasons to focus our attention to the implications of the present take on RTM, which, as the Complementarity Claim shows, has much more practical utility than RTM is typically credited for (Mari *et al* [2017]; Boumans [2016]; Reiss [2008]).

Moreover, the Complementarity Claim is valuable even if it does not constitute a direct objection to claims about the incompatibility of psychometrics and RTM. The diagnosis that underlies the Complementarity Claim clarifies the interpretations under which psychometrics and RTM *are* complementary. Thus the argument should help end confusion concerning psychometrics and RTM, not because it has been conclusively shown that RTM and psychometrics are complementary under all interpretations, but because the present analysis allows us to pinpoint differences in assumptions that lead to different conclusions about the compatibility of RTM and psychometrics. In addition, the Complementarity Claim calls us to question the present separation of the two approaches, which neglects their potential for joint usage.

Let me finish off with some remarks on why the fruits of complementarity should tempt proponents of both approaches. Start with psychometrics. Several authors have argued that psychometricians tend to assume that their target attribute can be measured on an interval scale, but this assumption is not scrutinized or established empirically (e.g. Borsboom and Zand Scholte [2008]; Michell [2008]; Hobart *et al* [2007]). Part of the problem is that Classical Test Theory is the dominant approach in psychometrics, but justifying interval scale properties with CTT is notoriously difficult (Embretson and Reise [2000], 28-32). Clearly, the interpretability of measurement results suffers if we do not know that an interval level interpretation of the measurement data is justified, not least because the choice of appropriate statistical tests depends on it. Incorporating RTM-based considerations in psychometric practice would allow for a wider range of hypotheses to gain genuine evidential support from psychometric instruments. Note however that even staunch supporters of RTM admit that psychometric tests can be powerful predictors in the absence of

Representational Interpretability (see for example Suppes and Zinnes [1963]). Lack of grounding in Representational Interpretability does not make a psychometric test useless, but it does undermine its status as measurement.

What about proponents of RTM? Economists are notably fond of proving representation and uniqueness theorems – the advancement of decision theory is a striking example of the allegiance between economics and RTM. But economists are frequently charged with oversimplifying the attributes they want to measure. In particular, the everyday notion of preferences as unobservable, mental desires has been impoverished into the much criticized 'preferences as choices' notion, arguably because capturing preferences qua mental entities is way more complex than tracking choice behavior (Angner [2013]). In so far as impoverished conceptualizations of the target attribute are unsatisfactory, economists would benefit from an increased understanding of how psychometricians handle the measurement of unobservable target attributes. In particular, psychometric tools for studying the properties of unobservable attributes show that economists cannot appeal to the impossibility of measuring unobservables when justifying the move from preferences qua desires to preferences qua choices. It remains to be seen what other fruits complementary usage of RTM and psychometrics can reap.

Acknowledgements

I am grateful to Anna Alexandrova and Hasok Chang for reading and commenting on multiple earlier drafts. I would also like to acknowledge help from the following people, who provided useful comments on or related to drafts of this paper: Juha Haaja, Conrad Heilmann, Mats Ingelström and Jacob Stegenga. I thank the following institutions for research funding: Arts and Humanities Research Council Doctoral Training Partnership Cambridge; the British Society for the Philosophy of Science; Cambridge European Trust; Newnham College Cambridge.

Elina Vessonen
Department of History and Philosophy of Science
University of Cambridge
Cambridge, United Kingdom
esmv2@cam.ac.uk

References

Alexandrova, A. [2017]: *A Philosophy for the Science of Well–Being,* Oxford: Oxford University Press.

Alexandrova, A., and Haybron, D. [2016]: 'Is Construct Validation Valid?', *Philosophy of Science,* **83(5)**, pp. 1098–1109.

Andrich, D. [1988]: *Rasch models for measurement,* Newbury Park, CA: Sage.

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

[2011]: 'Current Trends in Welfare Measurement', in J. B. Davis and D. Wade Hands *(eds)*, *The Elgar Companion to Recent Economic Methodology*, Northampton: Edward Elgar.

Bagby R.M., Ryder A.G., Schuller, D.R. and Marshall, M.B. [2004]: 'The Hamilton Depression Rating Scale: has the gold standard become a lead weight?' *American Journal of Psychiatry,* **161(12)**, pp. 2163–77.

Bond, T., and Fox, C. M. [2001]: *Applying the Rasch model: Fundamental measurement in the human sciences*, Mahwah, NJ: Lawrence Erlbaum.

Borsboom, D., and Mellenbergh, G. J. [2004]: 'Why Psychometrics is Not Pathological A Comment on Michell', *Theory & Psychology*, **14(1)**, pp. 105–120.

Borsboom, D. and Zand Scholten, A. [2008]: 'The Rasch model and conjoint measurement theory from the perspective of psychometrics', *Theory & Psychology,* **18(1)**, pp. 111–117.

Boumans, M. [2016]: 'Suppes's outlines of an empirical measurement theory', *Journal of Economic Methodology,* **23(3)**, pp. 305–315.

Campbell, D. T. and Fiske, D. W. [1959]: 'Convergent and discriminant validation by the multitrait–multimethod matrix', *Psychological bulletin,* **56(2)**, pp. 81–105.

Cartwright, N., Bradburn, N. and Fuller, J. [2016]: 'A Theory of Measurement', *Durham University: CHESS Working Paper,* No. 2016–07, <https://www.dur.ac.uk/resources/chess/CHESSK4UWP_2016_07_BradburnCartwrightFuller.pdf>

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

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

Cronbach, L. J. and Meehl, P. E. [1955]: 'Construct validity in psychological tests', *Psychological bulletin*, **52(4)**, pp. 281–302.

Embretson, S.E. and Reise, S.P. [2000]: *Item response theory for Psychologists*, Mahwah (NJ): Lawrence Elbaum Associates.

Engelhard, G. Jr. [2013]: *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*, New York: Routledge.

Frigerio, A., Giordani, A. and Mari, L. [2010]: 'Outline of a general model of measurement' *Synthese,* **175**, pp. 123–149.

Hacking, I. [1983]: *Representing and intervening,* Cambridge: Cambridge University Press.

Hambleton, R. K., Swaminathan, H. and Rogers, H.J. [1991]: *Fundamentals of item response theory*, Newbury Park, CA: Sage.

Heilmann, C. [2015]: 'A New Interpretation of the Representational Theory of Measurement', *Philosophy of Science,* **82(5)**, pp. 787–797.

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

Hood, S. B. [2013]: 'Psychological Measurement and Methodological Realism', *Erkenntnis,* **78(4)**, pp. 739–761.

Isaac, A. [2017]: 'Hubris to humility: Tonal volume and the fundamentality of psychophysical quantities', *Studies in History and Philosophy of Science Part A*, **65–66**, pp. 99–111.

[2013]: 'Objective Similarity and Mental Representation', *Australasian Journal of Philosophy,* **91(4)**, pp. 683–704.

Judd, C. and McClelland, G. [1998]: 'Measurement' in S. Fiske, G. Daniel & L. Gardner *(eds), Handbook of social psychology, 4th edition*, Boston: McGraw–Hill, pp. 180–232.

Keats, J. A. [1967]: 'Test theory', *Annual Review of Psychology*, **18(1)**, pp. 217–238.

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

Krantz, D. [1991]: 'From Indices to Mappings: The Representational Approach to measurement', in D. Brown and J.E. Smith *(eds)*, *Frontiers of Mathematical psychology. Essays in Honor of Clyde Coombs,* New York: Springer–Verlag, pp. 1–52.

Kyngdon, A. [2008]: 'The Rasch model from the perspective of the representational theory of measurement', *Theory & Psychology,* **18(1)**, pp. 89–109.

[2011]: 'Plausible measurement analogies to some psychometric models of test performance', *British Journal of Mathematical and Statistical Psychology,* **64**, pp. 478–497.

Luce, R. D. [1959]: ‘On the Possible Psychophysical Laws’, *Psychological Review*, **66**, pp. 81–95.

[1997]: ‘Several unresolved conceptual problems of mathematical psychology’, *Journal of mathematical psychology,* **41(1)**, pp. 79–87.

Luce, R. D., and Tukey, J. W. [1964]: ‘Simultaneous conjoint measurement: A new type of fundamental measurement’, *Journal of mathematical psychology,* **1(1)**, pp. 1–27.

Luce, R. D., Suppes, P., and Krantz, D. H. [1990]: *Foundations of measurement: representation, axiomatization, and invariance*, London: Academic Press.

Mari, L. [2000]: ‘Beyond the representational viewpoint: a new formalization of measurement’, *Measurement*, **27(2),** pp. 71–84.

Mari, L., Carbone, P., Giordani, A. and Petri, D. [2017]: ‘A structural interpretation of measurement and some related epistemological issues’, *Studies in History and Philosophy of Science Part A*, **65–66**, pp. 46–56.

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

McClimans, L., Browne, J. and Cano, S. [2017]: ‘Clinical outcome measurement: Models, theory, psychometrics and practice’, *Studies in History and Philosophy of Science Part A*, **65–66,** pp. 67–73.

Michell, J. [2014]: ‘The Rasch paradox, conjoint measurement, and psychometrics: Response to Humphry and Sijtsma’, *Theory & Psychology*, **24(1)**, pp. 111–123.

[2008]: ‘Is psychometrics pathological science?’, *Measurement*, **6(1–2)**, pp. 7–24.

[1993]: 'The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell', *Studies in History and Philosophy of Science Part A*, **24(2)**, pp. 185–206.

[1986]: 'Measurement scales and statistics: A clash of paradigms' *Psychological bulletin,* **100(3)**, pp. 398–407.

Narens, L. and Luce, R.D. [1993]: 'Further comments on the "nonrevolution" arising from axiomatic measurement theory', *Psychological Science,* **4(2)**, pp.127–130.

Perline, R., Wright, B. D. and Wainer, H. [1979]: 'The Rasch model as additive conjoint measurement', *Applied Psychological Measurement*, **3(2)**, pp. 237–255.

Rasch, G. [1960]: *Probabilistic models for some intelligence and attainment tests,* Copenhagen: Denmark Paedagogiske Institute.

Reiss, J. [2008]: 'Error in Economics: Towards a More Evidence–Based Methodology', London: Routledge.

Riordan, S. [2015]: 'The objectivity of scientific measures', *Studies in History and Philosophy of Science Part A,* **50**, pp. 38–47.

Soler, L., Wieber, F., Allamel–Raffin, C., Gangloff, J. L., Dufour, C., and Trizio, E. [2013]: 'Calibration: A conceptual framework applied to scientific practices which investigate natural phenomena by means of standardized instruments', *Journal for general philosophy of science*, **44(2)**, pp. 263–317.

Stevens, S. S. [1951]: 'Mathematics, Measurement, and Psychophysics', in S. S. Stevens *(ed.)*, *Handbook of Experimental Psychology,* New York: John Wiley, pp. 21–29.

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

Suppes, P. and Zinnes, J. L. [1963]: *Basic measurement theory,* in R.D. Luce, R.R. Bush and E. Galanter *(eds), Handbook of mathematical Psychology*, Oxford: Wiley, pp. 1–76.

Tal, E. [2016]: 'Making Time: A Study in the Epistemology of Measurement', *British Journal for the Philosophy of Science*, **67**, pp. 297–335.

[2015]: 'Measurement in Science', in E. Zalta *(ed.), Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/sum2015/entries/measurement–science/>.

[2013]: 'Old and New Problems in Philosophy of Measurement', *Philosophy Compass,* **8(12)**, pp. 1159–1173.

[2012]: *The Epistemology of Measurement: A Model–Based Account.* Doctoral Dissertation. University of Toronto.

Wright, B. D., and Stone, M. H. [1999]: *Measurement essentials.* Wilmington: Wide Range Inc.

van Fraassen, B. [1980]: *The Scientific Image*, Oxford: Oxford University Press.

[2008]: *Scientific Representation: Paradoxes of Perspective*. Oxford: Oxford University Press.

Vessonen, [2017]: 'Psychometrics versus Representational Theory of Measurement', *Philosophy of the Social Sciences,* **47(4–5)**, pp. 330–350.