Rating Scales From An RTM-perspective: Not What They Seem To Be

Maas van Steenbergen [1,*]

[1] *Faculty of Behavioural and Social Sciences, Methodology & Statistics, Utrecht University, the Netherlands*

Correspondence*:
Corresponding Author
m.vansteenbergen@uu.nl

> "*Eh bien!*"–exclaimed Walras characteristically–"this difficulty is not insurmountable. Let us suppose that this measure exists, and we shall be able to give an exact and mathematical account of it". [. . .] In view of the fact that theoretical science is a living organism, it would not be exaggerating to say that this attitude is tantamount to planning a fish hatchery in a moist flower bed".
> – *Nicholas Georgescu-Roegen*

> "I think the foundations of measurements [. . .] have a lot of implications about the way you do and actually think about measurement. There is a great deal of feedback from work on foundations on the actual practice, unlike a lot of other fields of mathematics, where work on foundations is divorced from the actual practice of other disciplines".
> – *Amon Tversky*

1 INTRODUCTION

Variables in psychology often aim to quantify strength of sensation and personality traits, or the level of attitudes and ailities. Because of the inherently subjective nature of these variables, we cannot measure or validate those concepts external to the participant. In other words, we cannot look in their heads and experience what they experience. This makes it more challenging to assess subjective variables and more challenging to assess the underlying structure. Psychologists, however, developed several tools to estimate the value of such variables. One of those techniques is the humble rating scale. This instrument involve a series of increasing numbers, often numbered from 1 to 5, 1 to 7, or 1 to 100. Participants then have to rate themselves on this scale. The different scores can be labelled. Sometimes all answering options have a label, while sometimes only the first and the last option are labelled. Sometimes they are balanced, with shifted numbers so that the middle option is 0. Participants then have to select one of the options that best describes their state of being on a construct of interest. Over the course of time, some researchers have become accustomed to treating these variables as continuous, while others claim that they are not, and you should use non-parametric methods (**?**). Parametric methods are only allowed after meeting some criteria. A popular prescription, for example, is that you can treat an ordinal variable as continuous as long as you have more than a certain number of answering options (**?**). Another popular prescription claims that ordinal variables become continuous after summing a sufficient number of them that measure the same construct. Conversely, a group of statisticians is firmly against this treatment. The conflict was never resolved,and neither treatment is considered definitive.

In the current paper, we intend to study how different assumptions about the mapping of the quantitative structure of a variable on a rating scale influence the number of false positives. We do this by imagining a perfect (ratio-level) continuous homomorphic mapping of a *quantifiable* psychological variable. Then, we specify a certain structural relationship between the perfectly measured variable and an ordinal or interval equivalent. We use a series of increasingly loose restrictions to the mapping, studying how the original variables are deformed when looser restrictions are taken. Then, we will administer some simple statistical tests to see the degree to which the conclusions based on the 'perfect' homomorphic mapping of the score correspond to ratings on the rating scale. We use this to argue that one clearly needs to define how the test relates to a hypothetical zero-point and to the stability of the mapping. What is the influence of stability, the scaling, and the zero point for psychological tests? On the other hand, we argue that this is a question of *degree*. We show that stronger or weaker violations of these assumtpions are possible. This can be expressed relatively, if we base the judgment of i.

Our choice to give the hypothetized perfect measurement a perfect zero point is based on several observations. First, the absence of an observable zero-point or the claim that one never really lacks happiness fully are no challenge for the claim that there is no zero-point at all. Temperature, for example, does have an absolute zero-point, at 0° K, yet it cannot ever be reached. Nobody would argue because it cannot ever be reached that we should stop attempting to measure temperature in Kelvin. Using the absolute zero-point (or the absence of the posited attribute) also allows for the specification of an unambiguous

A reason for the current realist treatment is the frustration about the limitations of *modern* treatments of this topic. Many modern simulation studies implicitly assume the thing that ought to be proven: that the mapping of the continuous variable to the measurement scale has equal intervals.

For the current project, we focus our attention on cases where an attribute of interest is pricipially continuously measurable. Therefore, a continuous, homomorphic mapping of the qualitative structure towards a quantitative representation can be constructed. While there might be practical limitations in the construction of such a mapping, this is of no consequence to our experimental set-up: we posit to know the true value, and use this as a starting point for a simulation study.

1.1 Representational Theory of Measurement

We assume, in alignment with Krantz, Luce, Suppes, and Tversky (henceforth referred to as KLST), that an attribute can be quantified only if it is proven to be a quantity (**?**). We will use the criteria formalized in the representational theory of measurement (RTM). KLST is at present the most fully developed account of RTM. In line with this theory, we see measurement from the perspective of a representation theorem and a uniqueness theorem. The representation theorem asserts that a set, together with one or more relations on that set, follows certain axioms so that it can truthfully preserve the structure of qualitative attributes of an object to a numerical quantity. In other words, a *homomorphic* mapping of the qualitative relational structure into a quantitative representation can be constructed.

Suppose that the transformation function $\phi(a)$ is a homomorphism and the representation theorem holds for the mapping of qualitative ordering $\succ$ to quantitative ordering $>$. If $\phi(a)$ then maps a qualitative attribute of set A into $\mathbb{R}$, it should maintain the structure of the qualitative attribute in its numerical representation: if and only if a is qualitatively greater than b, it is numerically greater than b. The uniqueness theorem specifies the permissible homomorphic transformations of A that lead to the same structure of numerical relations. E.g., both Fahrenheit or Celsius measurements are legitimate homomorphic mappings that maintain the qualitative structure of temperature, differing in zero-point and scale. Note that these aspects of measurement do not concern modelling or statistical analysis, but are a *necessary condition* for those.

We will assume a version called "representational minimalism" that rids it of most of its epistemological basis so that it can serve as a common ground for discussion about measurement, recognizing that most of the critical literature about measurement takes place either within or in discussion of this framework (**?**). The acceptance of RTM is not universal and we do not pretend to be. We do not make firm claims whether a realist, an operationalist, or a representationalist epistemology should be matched to the formal aspects of the framework. We do, however, choose to use representationalist language, as this is used in the original representation of RTM and is the most well-known and fully realized. It should be noted that choosing a different framework means that the formalism might change. If this change is subtle, then the current study might still be interpretable. In some frameworks, however, especially those that doubt the quantifiability of psychological variables, the current experiment might be uninterpretable.

1.1.1 Ordinal measurement

An ordinal score would imply that the score is only dependent on rank order, and does not support concatenation operations. A score in P can only indicate that a score is higher or lower than another score in P. Based on the measure only, beyond that ordering, we know nothing about P. In other words, any mapping of which we can only assume that these characteristics hold is an ordinal mapping of the relational structure of Q. Note that an ordinal mapping of the structure of Q can be made if a ratio-mapping can also be made, but not vice versa.

This can also be written down axiomatically. Let X, Y & Z by any strongly tied elements for Q. Then the projection of the quantity to an ordinal scale, which we will refer to as P, holds to the following conditions. Let A be a set and $\succeq$ be a binary relation on A.

- if $X \succeq Y$ & $Y \succeq Z$, then $X \succeq Z$ (transitivity);
- Either $X \succeq Y \parallel Y \succeq X$ (connectedness);
- If $X \succeq Y$ & $Y \succeq X$, then $X = Z$ (antisymmetry).

Two elements have a weak order iff, for all a, b, axiom 1 and 2 are met. If the third axiom is also met, then the ordering is strong and elements cannot be tied. If these axioms for the empirical

1.1.2 Ratio measurement

As for the 'perfect' ratio-mapping of the principially measurable attribute Q we will assume that the following additional characteristics will hold (**?**):

- $X \oplus (Y \oplus Z) = (X \oplus Y) \oplus Z$ (associativity);
- $X \oplus Y = Y \oplus X$ (commutativity);
- $X \succeq Y$ iff $X \oplus Z \succeq Y \oplus Z$ (monotonicity);
- if $X \succ Y$ then there exists a Z such that $X = Y \oplus Z$ (solvability);
- $X \oplus Y \succ X$ (positivity);
- there exists a number n such that $nX \succeq Y$ (where $1X = 1$ and $(n \oplus 1)X = nX \oplus X$) (Archimeadean condition).

This essentially means that a value q can always be put in terms of another value r in Q. Every ratio-scale is homomorphic to the qualitative attribute that is being measured, and ratio-scales are thus homomorphic to each other (**?**). The last axiom (the archimedean condition) is added to ensure that the set of possible scores is finite. E.g., say that we have developed a standard measure for happiness: H. Then we can say that any value x is written in terms of H iff the finite ratio $\frac{x}{HCS}$ holds[1].

1.1.3 Rating scales and ordinality

Likert scale ratings are normally understood to be ordinal, but we argue that this is not necessarily the case within the context of an experiment. If the ordering of attributes or their relationship to the hypothetical perfect measure is not invariant to time and is not invariant between people, then the relationship of the attribute to the rating-scale representation of that attribute is not consistently weakly ordered. To show this, we first need to formalize the connection between the hypothetical perfect measure and a measure that is ordinal for one specific person at one specific time.

Assume that the empirical structure of an attribute is in correspondence with the axioma's for ratio measurement scaling. This means, by extension, that all axioma's for ordinal measurement are in correspondence with the empirical structure of that attribute. Let us assume we have a hypothetical homomorphic mapping of the empirical structure of this attribute expressed quantitatively (with ratio-scaling). Then, each possible weakly[2] ordinal representation of quantitative values of this perfect measurement measure can be represent to provide an unambiguous weakly ordered classification of scores in Q to scores in terms of P[3] Assume that instrument P has n elements. Then we can define a series of $n-1$ 'thresholds' R as $\{a, b, c \dots z\}$. These thresholds are the point of q where a score p_1 in P jumps to another classification p_2 in P on the basis of a score q in Q. We then define a score p in P for each q in Q as follows, assuming we have an n-level measurement scale for P:

$$\begin{cases} 1 & q \leq a \\ 2 & a \leq q < b \\ 3 & b \leq q < c \\ \dots & \dots \\ n & z < q \end{cases} .$$

Note that if the mapping does not reflect these conditions, then the first axiom for ordinality is broken. Thus, this step function is a logical consequence of the empirical structure of the variable. It needs to be met. Otherwise, the two representations of the attribute are mutually inconsistent. Therefore, this representation follows from the assumed structure of our hypothetical perfect representation of the variable and the ordinal rating scale.

While a strong ordering of an ordinal representation of the variable is a unique representation of the variable up to monotonic transformations, a representation that groups values into a certain number of equal segments can be incosistent with the same representation. In fact, many different constructions can

[1] We agree with the point by Franz that it is misleading to give physical magnitudes as examples when discussing psychological phenomena (**?**). Therefore, we choose to use psychological examples when we are thinking about psychological phenomena. This adds an extra face validity check: does discussing psychological attributes in this manner make sense?

[2] Meaning elements can be tied, see above.

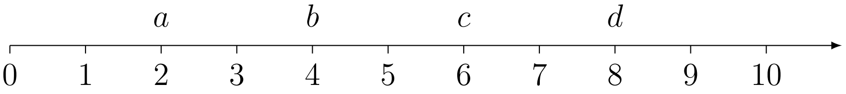
[3] This lack of ambiguity is only a result of this formalization. We may find that the relationship of the 'real' variable to its ordinal estimate is not quite so straightforward. We will discuss this later on.

be made. Say we represent an element ordinally, and there are no equivalent items. We can construct an ordering of a set of elements 1, 2, 3, …, 99, 100 of our continuous quantitative variable and we would like to map this to our ordinal rating scale. Then both

$$\begin{cases} 1 & q \leq 22 \\ 2 & 22 \leq q < 48 \\ 3 & 48 \leq q < 59 \\ 4 & 59 \leq q < 82 \\ 5 & 82 < q \end{cases} \& \begin{cases} 1 & q \leq 20 \\ 2 & 20 \leq q < 40 \\ 3 & 40 \leq q < 60 \\ 4 & 60 \leq q < 80 \\ 5 & 80 < q \end{cases} .$$

are valid representations of the variable with 5 equivalence sets. Yet, they are inconsistent to each other, because one number can be sorted into one category that would be sorted into another category in the other representation. E.g., the value 21 is seen as a 1 in the first representation while a value of 21 is seen as a 2 in the second representation.

This relationship can also be visualized on a number line. Consider a five-point instrument is used to measure a quantitative psychological variable. At a particular point in time t, we assume that the projection on the real number line is divided into four segments of equal length and one element which goes on infinitely. It should be noted that the scaling is arbitrary, but it has been set here to multiples of two for convenience.



The relationship between P and Q in this case can then be described with the following step function:

$$\begin{cases} 1 & q \leq 2 \\ 2 & 2 \leq q < 4 \\ 3 & 4 \leq q < 6 \\ 4 & 6 \leq q < 8 \\ 5 & 8 < q \end{cases} .$$

The spacing between two thresholds is once again set to two. Because the spacing between pegs is consistent over the distance, the measuring instrument . We note that we measure it measures at interval scale if the range of possible observations is limited to . If ceiling or floor effects are present,

The instability can either

1.2 Measurement instability

Measurement instability

1.2.1 Zero-point instability

Zero-point instability means that the zero-point of the mapping is unstable. The absolute zero-point of the real quantity is by definition stable, so it means that the entire mapping moves relative to the absolute zero point of the real quantity by a constant amount in the same direction.

1.2.2 Scaling instability

Scaling instability is instability that comes from a changing distance between the highest threshold and the lowest threshold (the zero-point). The distance of the range of the measurement instrument is unstable. In other words, the segment of the real quantity that is mapped to the representation can be wider or more narrow in the real quantity depending on the person or the timepoint on which a person is measured. Scaling instability is independent of zero-point instability. The range of the real variable is not affected by the scaling instability of the mapping.

1.2.3 Interval instability

Interval instability means that the distance between thresholds aside from the last and the first threshold is unstable. If intervals are unstable, the distance between each consecutive pair of thresholds is not identical. Interval instability can have a structural or a random appearence. Structural An example of structural interval instability would be a growing distance between each subsequent consecutive pair of thresholds. In random structural interval instability, there is no relationship between the order of the pairs and Interval instability can simultaneously have a structural or a random component. person-to-person and from timepoint-to-timepoint.

1.2.4 Combinations

These three instabilities operate mostly independently but a few things should be noted. First, if the zero-point of the mapping is stable then scaling instability can only affect the right bound of the range. Further, thresholds cannot be outside of the range of the measuring device by definition. Therefore, interval instability is limited in range through scaling and zero-point instability. We also reiterate once more that the values of the actually existing quantity are not affected by the mapping. We consider these absolute up to the point of the uniqueness theorem, which is the scaling. The chosen scaling of the real quantity is of no concern. If the original attribute would be scaled differently, it would affect the thresholds proportionately. Statistical conclusions would not be impacted. All forms of instability can be either systematic or unsystematic. Systematism can be consequential in the context of applying a statistical test. If the instability is systematic, then the mapping of the variable from an existing quantity to the rating score is impacted by the value of one or more independent variable. If the instability is unsystematic, there is no relationship between the

1.3 Between-person variation

It is often assumed that aggregations of individual measures rescind the limitations of different scaling because these measurement differences cancel out when enough samples are drawn. The origins of this assertion are from Knapp. This idea can be seen as follows: if the limit of the number of bins becomes infinite, and the underlying variable is continuous, then the measuring instrument will approach the normal distribution.

1.4 Within-person variation

When studying within-person mappings, we can assume that those psychological processes part of Q are time-dependent. These measures are shaped by different forces and the previous states of the construct (?). Their values are continuous and are related to each other in a structured manner (?). We also assume that their values are differentiable over time, changing smoothly. Their values can increase or increase very quickly, but not instantaneously. This, they are modeled best using the rate of change of the variable, through differential equations (?).

Making a, b, c, and d functions of time, step function F(x, t) becomes:

$$\begin{cases} 1 & q \leq a(t) \\ 2 & a(t) \leq q \leq b(t) \\ 3 & b(t) \leq q \leq c(t) \\ \dots & \dots \\ n & z(t) \leq q \end{cases}.$$

This means that each threshold becomes a function, with time as input. That is, for each time t, a .

1.5 Using the framework for simulation studies

If you want to make claims about the consequences of different structural deficiencies of a framework, it can be helpful to run a simulation of the effect of those deficiencies to show the consequences of those effects to researchers.

1.6

We used the Julia language, and in particular the ‘Statistics.jl’ packages, to set up the simulation and to run the statistical analysis (? ? ?). Analyses were run on a personal computer. Full information about dependencies and version numbers can be found in a machine-readable format in the Manifest.toml file in the Github-repository. Instructions for running the analysis through a sandboxed project environment identical to our system can be found on the main page of this repository.

1.7

Both the distributions and the test were chosen to be the simplest test possible to show the potential consequences of zero-point, scaling, and interval instability. The experiment can be repeated using any possible test. If the reader decides that we are in error. To disarm at least one possible counter, we also perform a signed rank

1.8 Step 1: Specifying the parameters

Initially, we specified a statistical test so that we would be close to a statistical power of 0.80 based on an α of 0.05. These are greatly valued, standard. We used standardized normal distributions, as the scaling is arbitrary as per the uniqueness theorem. For each simulation, we had two situations. One in which we sampled two times from a standard normal distribution. In another test, we used one sample with a μ of 1 and a σ of 1 and one sample with a μ of 0 and a σ of 1. We have chosen a sample size so that our reasoning is consistent a power of 0.80.

1.9 Step 2: Mapping the parameters to ordinal values

We developed a mapping function that changes the thresholds depending on the characteristics of the .

1.10 Step 3:

2 DISCUSSION

and you are bound to get similar results. Why do a simulation study, then? To express a number of points. The first point is that scales require validation in the broadest sense of the word: it should measure what it purports to measure. Measuring what it purports to measure can be seen to mean that the quantity of something that is there has a bearing on the measuring instrument that we are using.

Also, this study needs to be seen as an attempt at a bridge between psychometrical testing theory and formal measurement theory. These two groups usually communicate somewhat adverserially, yet seem to be unable to consider the . The scope of a fully weighed-up

2.1 Why is scale validation so challenging?

Concatenation of lengths is an almost

2.2 Representationalist and realist

3

3.1 Relationship to Classical Test Theory

In classical test theory,

3.2 Relationship to Latent Variable Theory

3.3 Relationship to the Rasch Model

3.4 Validity

We subscribe to the point of . However, we do not agree that it should be so casually treated as a question that can be answered with a simple yes or no. Not only the entity

3.5 Measurability

The experiment above aims to show that the factors that go into scale instability, discretization, can influence.

Some concepts are easier to measure than others.

In the social sciences, there is often something left when you try to quantify kitchen-and-sink concepts that is not captured in the measurement (). The economist Georgescu-Roegen has called this the qualitative residual (**?**). It is not far-fetched to assume that this qualitative residual plays a role in the difficulty of achieving scale stabiblity or defining the relative zero-point. Georgescu-Roegen asserts the importance of quantifying that residual regardless of the challenges: ‘We buy and sell land by acres, because land is often homogeneous over large areas; and because this homogeneity is not general, we have differential rent. How unimaginably complicated economic life would be if we adopted an ordinal measure of land chosen so as to eliminate differential rent, let alone applying the same idea to all economic variables involving qualitative variations’ (**?**)!

With our results, we aimed to show that imperfect data can lead to correct results in many cases, and that a deviation from an ideal does not, by itself, mean that a study is pointless. In other words, we agree that the methods are imperfect and that bad measurement instruments can have terrible consequences for the validity of a research project. Yet right now they allow for a useful abstraction to study variation in populations in ways that are difficult to express qualitatively (). If we want to study these concepts properly, we need to use these methods productively *and* we need to be able to express their limitations. That is why it is a shame that the people who use the methods and the people who critique inhabit different worlds.

What is problematic, however, is that we do not know how our variables relate to

That is not to say that no steps should be taken to actively improving the way measurement is used in the social sciences. Much beyond the confines of what statisticians and psychometricians might be comfortable with. We do not intend to sideline the discussion of the consequences of the imperfection of scales. Quite the opposite. Just consider the replication crisis, the lack of epistemic iteration, the sometimes grating examples of methodolatry. By themselves, any of these discussions could warrant improving our understanding of where scales fall short. But too often, the conversation is had adverserially. The broad historical insight of many critics could be put to better use.

If there is no one-to-one relationship from the measurement device to the measure, it is wise not to mistake the measurement scores for the empirical relationships that are described. This problem is called the ‘map-territory’ problem in the philosophy of science. Rating scales were invented to make an estimate of the level of psychological variables in certain contexts, but they make a whole lot of assumptions about the nature of what those variables are. Current r Those ass

Going o

It must also be said that we do feel some sympathy towards the viewpoint by Sijtsma(**?**).

We agree with Borsboom, Mellenbergh and van Heerden that what validity should be defined as a scale measuring what it purports to measure. However, their conception is probabl. A scale might very well be able to measure what it purports to measure, but if this is only an attempt

3.6 Utility Measures

An hereto unnoticed (as far as I know) strong resemblance to the discussion from economists about the status of utility measures. This problem is in many ways identical to the problems that we see in psychology. What are the limits of

There is a distinction between unobservable and observable properties made in psychology that pops up often, such as in the discussion of latent variables. This distinction is not without criticism. As Burgos notes, the distinction is a (**?**)

3.7 Commonalities and Differences Between Our Approach and Latent Variable Theory

3.8 Practical Recommendations

* Don’t follow the regular guidelines * The *operational ideal* should be to minimize the deviation between a real measure and the obtained measure (**?**). This can be through improved scaling, more developed non-parametric methods . If a measure is imperfect, it should be admitted, and used in uncertainty calculations. For this, it is necessary

3.9 Recommendations

1. Think

NOTATION

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

FUNDING

No external funding was used for this project.

ACKNOWLEDGMENTS

I acknowledge the work of my thesis supervisors, who introduced me to the method and left me free to pursue the project as I imagined it in all its weird and shifting shapes. The great help of the Julia community was also appreciated, as they have been pushing me forward where I got stuck and took the time to respond to my stupid questions. Finally, I would like to acknowledge the feedback and conversations between me and my thesis group, who have been working through my text and made sure that it is easy to follow and well-written. Special thanks to Giuliana Orizzonte, Daniel Anadria, dr. Hessen, dr. Derksen, dr. Grelli and dr. Bringmann for fruitful discussions and feedback about my topic. Lastly, I would like to thank my girlfriend, family, and friends for the mental support throughout.

DATA AVAILABILITY STATEMENT

The code, additional material, and generated data for this study can be found on GitHub.