frontiers

Statistical Consequences of Mapping Instability from Rating Scale Scores

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 practice of other disciplines".
> – *Amon Tversky*

1 INTRODUCTION

Variables in psychology often aim to quantify strength of sensation, personality traits, or attitudes. Because such variables are based on the experience of people of their subjective feelings and opinions, we cannot measure or validate those concepts without participants conveying them to researchers. In other words, we cannot look in participants' heads and experience what they experience. Participants need to communicate their 'state' of a variable of interest to us. Within psychological science, these are often thought of as quantitative or ordinal magnitudes. Many tools are invented that aim to estimate the value of variables that are amenable to measurement. The rating scale is the most popular family of instruments by far. This family of instruments always involves a series of ordered options that represent an increasing magnitude on a variable of interest. Its strengths include its low price of administration, the easy reproducability, and its willingness to appear similar in different modalities. Many variants are popular: linear numeric scales, likert scales, frequency scales, and the visual analog scale. There is a great variety in the details of the labelling, the number of categories, and in the way that the answer categories are presented graphically.

Yet, it can be unclear what values generated using rating scales mean. Scoring 'extremely likely', '4' or 'very often' on a rating instrument does not have the same unambiguous meaning as a read-out of 5.2 cm on a piece of measuring tape. There is a clear contrast between estimates of psychological variables with the measures of length, of temperature, or of weight. When we express empirical relationships in measures

with a commonly known scaling, we know exactly what quantity the measure indicates, what the margin of error can be, and how this error relates to the unit .

This difference has traditionally been understood as the difference between ordinal scales on the one hand, and interval- or ratio-level scales on the other hand. This distinction was introduced by Stevens. When an attribute is scaled ordinally, we know that the scores can only say something about relative ordering: we know that 'very likely' is higher than 'likely'. With interval- and ratio-level scales, the empirical relationships are expressed in terms of the ratio of the attribute of one object to a standard score of an attribute: the unit. Stevens saw the type of measurement as reflective of empirical relationships of attributes present between objects. He developed this theory, the Representational Theory of Measurement, in response to criticism by physicists of measurement in psychology that did not adhere to a classical conception. In the classical conception of measurement, only quantitative-level measurement, was considered meaningful.

However, by making an additional assumption we aim to show that aggregations of rating scale measurements are not ordinal in the traditional sense. This is even when the underlying estimated magnitude is principally amenable to quantitative or ordinal measurement. By explicitly assuming that the structure of the empirical relationships in nature are quantitative, which is often already assumed, we can show that rating estimates can lead to incorrect statistical results even when they are treated as ordinal. By postulating that we know the magnitude of a quantity that the rating scale indicate, we can map these quantities to rating scale levels which are ordinal in nature. In the appendix, axioms for ordinality and , based upon the standard work in the field, are included.

We intend to study how different assumptions about the mapping of the quantitative structure of a variable on a rating scale influence the power and the Type I error rate. 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 equivalent with a set number of *equivalence classes*: items with the same score. If this structural relationship changes between participants, we say that there is *measurement instability*. We use a series of increasingly loose restrictions to the mapping, studying how the original variables are deformed when looser restrictions are assumed. 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. We also study the influence and make a case for the importance of stability, scaling, and zero point for psychological rating scales. Finally, we make clear that this is necessarily a question of *degree*. We show that stronger or weaker violations of these assumptions are possible.

For the current project, we focus our attention on cases where an attribute of interest is pricipially continuously measurable. Therefore, a *homomorphic* mapping of the qualitative structure can be constructed. This mapping is a quantity: a ratio-level measurement with an absolute zero-point. We posit to know the 'true value' of this attribute. We assume that the quantity of the attribute maps to the rating scale item in question. We map this quantitative representation again for each individual rating scale representation, adding some difference. Then, we apply the t-test to all of these different mappings. We call the differences in mapping *measurement instability* and make a distinction between four kinds: zero-point instability, scaling instability, and systematic and random threshold instability. *Zero point instability* means that the relative zero-point changes between participants. *Scaling instability* means that the distance between the zero-point and the maximum changes. *Systematic threshold instability* means that there is a patterned . and test the robustness of these measures. The goal of this simulation framework is to be able to study the characteristics

1.0.1 Rating scales and Ordinality

While aggregations of rating scale responses are normally understood to be ordinal, we argue that this is not true if there is between-person or within-person scaling instability. 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 axioms for ratio-level measurement scaling. This means, by extension, that all axioms 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[1] ordinal representation of quantitative values of this perfect measurement measure can result in an unambiguous weakly ordered classification of scores in Q to scores in terms of P[2] Assume that instrument P has n elements. Then we can define a series of $n-1$ 'thresholds' T as $\{a, b, c \ldots z\}$. 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:

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

Note that if the mapping is inconsistent with this formalism, then the first axiom of 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. This representation follows from the assumed structure of our hypothetical perfect representation of the variable and the assumed structure of an ordinal rating scale response.

An ordinal representation of a continuous that groups values into a certain number of equal segments can be inconsistent with another representation of that variable that splits the range of the original variable into an identical number of equivalence sets. In fact, many different constructions can be made that are not compatible with each other. Say we want to map a continuous variable into five categories. The values of our continuous variable fall in range [0, 100] and we would like to map this range to five equivalence sets.

If so, then both

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

[2] 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.

$$p_1 = \begin{cases} 1 & q \leq 22.4 \\ 2 & 22.4 \leq q < 48.9 \\ 3 & 48.9 \leq q < 59.4 \\ 4 & 59.4 \leq q < 82.2 \\ 5 & 82.2 < q \end{cases} \quad \text{and} \quad p_2 = \begin{cases} 1 & q \leq 20.0 \\ 2 & 20.0 \leq q < 40.0 \\ 3 & 40.0 \leq q < 60.0 \\ 4 & 60.0 \leq q < 80.0 \\ 5 & 80.0 < q \end{cases}.$$

are valid representations of the variable with 5 equivalence sets. Yet, they are inconsistent, because a quantity can be sent by the mapping to one category that would be sent to another category in another mapping. E.g., the value 21.1 is seen as a 1 in the first representation while a value of 21 is seen as a 2 in the second representation. This breaks the axiom of transitivity. If we were to sum scores gained through multiple ordinal representations of the same score, the mapping becomes inconsistent between these representations. Note also that this is not equivalent to measurement error per sé, but relates to *measurement instability*. The issues come to be through the incompatibility of two representations of empirical relations.

1.1 Between-person mapping inconsistencies

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.

1.2 Within-person mapping inconsistencies

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 (**?**). We assume that the mapping function of a variable develops in the same way: continuously, changing smoothly, and differentiable over time.

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.

1.3 Measurement instability

We introduce the *measurement instability framework* to study the consequences of measuring procedures .*Measurement instability* can be used to make a distinction between the influence of different parameters that affect the scaling of the mapping function. When we introduce these topics, we assume that the other measurement instability parameters are stable. We also assume equal intervals, unless otherwise specified. Note that none of the measurement instability parameters influence the quantity of the underlying variable, it only has repercussions on the mapping of the quantity to the measure.

1.3.1 Zero-point instability

Zero-point instability means that the zero-point of the mapping is unstable. Because the absolute zero-point of the real quantity is stable, the thresholds move away or towards the absolute zero point of the real quantity by a constant amount in the same direction. The range of the thresholds will stay identical. Every threshold shifts by a constant amount. Therefore, assuming we have an n-level measurement scale for p:

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

where set of thresholds $T = \{a, b, c, \dots, z\}$, C_i is the constant for person or moment i, and q is a continuous measure of Q.

1.3.2 Scaling instability

Scaling instability is instability that comes from a changing distance between the highest threshold and the lowest threshold (the zero-point). The 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 difference of each threshold to the lower range bound is scaled up by a common factor. Therefore, assuming we have an n-level measurement scale p:

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

where where set of thresholds $T = \{a, b, c, \dots, z\}$, x_i is the scaling factor, and q is a continuous measure of Q.

1.3.3 Interval instability

Interval instability means that the distance between consecutive pairs of thresholds is not identical. Interval instability can have a structural or a random appearence. An example of structural interval

instability would be a growing distance between each subsequent neighbouring pair of thresholds. In random interval instability, there is no relationship between the order of the pairs and the distance of neighbouring pairs of thresholds. Interval instability can simultaneously have a structural and a random component.

$$\begin{cases} 1 & q \leq (a + C_{i1}) \\ 2 & (a + C_{i1}) \leq q < (b + C_{i2}) \\ 3 & (b + C_{i2}) \leq q < (c + C_{i3}) \\ \dots & \dots \leq \dots \\ n & (z + C_{in}) < q \end{cases} .$$

where the set of thresholds $T = \{a, b, c, \dots, z\}$

1.3.4 Combinations

These three instabilities operate mostly independently but a few things should be noted. One is that real scores can fall outside of the range of the thresholds. These scores will naturally be mapped to the lowest or highest option. Second, 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 by 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. Moreover, the chosen scaling of the real quantity is of no practical concern. If the original attribute would be scaled differently, it would affect the thresholds proportionately. Statistical conclusions would be identical after these scaling differences are propagated. Lastly, all forms of instability can be either systematic or unsystematic. If the instability is systematic, then the mapping of the variable from an existing quantity to the rating score is impacted by other variables of relevance to the test. If the instability is unsystematic, there is no relationship between other relevant variables and the mapping. Systematism can be consequential in the context of applying a statistical test, because if the mapping is inconsistent between test conditions it might lead to biased results.

2 METHOD

2.1 Aims

The aim of this simulation is to illustrate and clarify the potential consequences of measurement instability. Given these aims, we initially have decided to use the simplest test and data-generating mechanisms we could think of.

2.2 Data-generating mechanism

Our data-generation pipeline consists of multiple stages. First, we generate the data on the data itself.

We generate data for a simple independent-sample t-test with the aim that this test performed on the unprocessed data has a statistical power of 0.80 for a one standard deviation difference based on an α of 0.05. Our sample size was chosen to ensure an empirical power of 0.807: 18 participants per group.

In the first test, the two distributions should be equivalent. Two samples are drawn from the same distribution. Therefore, our test will be rejected in 0.05 percent of cases, as should be expected. We draw both samples from a standardized normal distribution:

$$\text{Normal } (\mu = 0; \sigma = 1)$$

In the second test, there is a one standard deviation difference between the null distribution and the alternative distribution. The parameters are as follows:

$$0\text{: Normal } (\mu = 0; \sigma = 1)$$
$$\text{A: Normal } (\mu = 1; \sigma = 1)$$

Random numbers are generated using the Xoshiro256++ algorithm. The seed is 8508845.

2.3 Threshold transformation pipeline

We have implemented a full-factorial design by planning a range of parameters for each sub-type of measurement instability. These factors include the distribution of the zero-point, the number of bins the data should be divided into, the distribution of the scaling, and structural threshold instabilty. We emulate differences in the degree of measurement stability by running the thresholds through a transformation pipeline. Each stage of the pipeline draws on the results of the last, until one set of thresholds is left. Thresholds are generated for each iteration of the simulation, right before we perform the analysis.

2.3.1 Generating the zero point

First we set a zero-point. This is based on a three standard deviation difference from the mean. If it is stable, the left bound is constant. If it is unstable, we use a normal distribution to generate a zero-point with μ is three standard deviations lower than μ is variable. In the last stage, this actual bound is cut off because the left-most category extends theoretically to lower values up to the absolute zero-point. Numbers are either a constant or drawn from a distribution. The constant is always -3. If drawn from a distribution, μ is always -3. σ is set at each interval when incrementing by 0.1 from 0.1 to 1.0.

2.3.2 Generating the scaling

We placed the maximum at the right bound of the 99% confidence interval on the basis of the null-data so that the mapping is equivalent for the two groups. The generation mechanism is similar to that of the zero-point: it is either a constant, or we use a normal distribution to generate a max-point where μ is equal to three standard deviations to the right of μ. The constant is thus always 6. If drawn from a distribution, μ is always 6. σ is set at each interval when incrementing by 0.1 from 0.1 to 1.0.

2.3.3 Setting the number of response categories

We set the number of response categories as one of the parameters for the simulation. This is fed into the function that generates the thresholds. The set of values includes each integer from four to fifteen, 20, 50, 80, and 100.

2.3.4 Generating the thresholds

We generated the intervals (all values of the thresholds aside from the outer bounds) by dividing the distance of the left bound of the range to the right bound of the range by the number of thresholds. Initially, this is always done

Unequal thresholds were generated using the cumulative distribution function of the beta-distribution. We chose this set-up because it results in either increasing or decreasing distance between, controllable by the β-parameter. If it is stable, the same parameters are chosen for each . If it is unstable, α and β were

2.4 Mapping the real scores to the bins

Afterwards, we cut the data using the thresholds we generated. This works using the simple step function described in the introduction: each value we have generated is sent to a value in the .

2.5 Analysis

We ran two independent two-sample t-tests to perform the analysis. We save the number

2.6 Software

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. Both the distributions and the test were chosen to use simple, commonly known statistics with the aim to show the potential consequences of zero-point, scaling, and interval instability. The experiment can be repeated using any possible test: the values are mapped *before* the analysis is run, so the full study is independent of the implementation of the test.

3 DISCUSSION

This study needs to be seen as an attempt at a bridge between scientists who study psychometrical testing, theoretical psychologists with an interest in formal measurement theory, and theoretical . We believe that the dots yell out to be connected, but it can be difficult to see the structural equivalence in . Either the abstract philosophical nature of

Another reason for the current 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 scales result in equal intervals. We believe that part of the reason for this confusion is that psychologists are allergic to making claims about the quantitavity of variables. Yet, most of the statistical tests we commonly use to test data rely on that claim. It can be difficult to understand why we rely on those assumptions for testing, but are reluctant when it comes to justification. This is an attempt to show that it can bring a great deal of clarity if we extend our lenience to allow for these assumptions in justification.

Our choice to give the hypothetized perfect measurement an absolute zero is based on both a defence and a justification. The justification for using the absolute zero-point (or the absence of the posited attribute) is the theoretical simplicity that is brought by an unambiguous relationship of the perfect quantity measurement and its relationship to a rating scale score. The defence is a response to the common claim that one never can never really observe the absence of a value of a trait. Each quantity has to have a theoretical zero-point, because otherwise the whole idea of quantity is meaningless. Temperature, for example, does have an absolute zero-point, at 0° K, yet it cannot ever be reached. Nobody would argue that because temperature is it cannot ever be reached that we should stop attempting to measure temperature in Kelvin.

3.1

The better anchored a psychological variable is to known

3.2 Ceiling Floor effect

3.3 Distributions: not as expected

3.4 Operational ideal

3.5 Relationship of this framework to other theories

We would like to

3.6 Relationship to Classical Test Theory

In classical test theory,

3.7 Relationship to Latent Variable Theory and Measurement Invariance

The current framework has the strongest resemblance to measurement invariance models Simulation studies comparing measurement invariance approaches

3.8 Relationship to the Rasch Model

3.9 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

3.10 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' (**?**)!

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

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. Mistaking the scores on that rating scale for the quantity of .

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

We align ourselves with "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. 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 the most well-known and fully realized. It should be noted that choosing a different framework means that the formalism might either change slightly or majorly. If this change is subtle, then the current study might still be interpretable. In some frameworks, however, especially those that doubt the hypothetical quantifiability of psychological constructs, the current discussion might be uninterpretable.

3.11 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 (**?**)

3.12 Takeaway points

1. When thinking about measurement, it can be useful to separate *measurement error* from *measurement instability*. The former is an error or deviation from a measure after already assuming its scaling. The latter is caused by inconsistencies in the mapping process from the real value to its rating scale equivalents.
2. The *operational ideal* of the measurement instability framework is to minimize the issues that inconsistent mapping between participants or timepoints can cause. This, for example, can be through rescaling or the development of statistical measures. Moreover, uncertainty calculations can be made on the basis of estimated measurement instability.
3. If it were easy to measure the relationship of the scaling of a rating scale and a 'real' magnitude of a quantity, it would have been as common as

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.

APPENDIX

3.13 Representational Theory of Measurement

The representational theory of measurement (RTM) is an important . While first formalized by Stevens, the most succesful

We assume, in alignment with Krantz, Luce, Suppes, and Tversky (henceforth referred to as KLST), that an attribute can be quantified only if its structural characteristics allow for measurement (). .In line with this theory, we see the measurability of an object from the perspective of a representation 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 preserves the structure of a qualitative attribute of an object to a numerical representation of that quality. In other words, a *homomorphic* mapping of the qualitative relational structure into a quantitative representation can be constructed. A uniqueness theorem specifies the set of allowable transformations which maintain this qualitative relation strcture. E.g., both Fahrenheit or Celsius measurements are legitimate homomorphic mappings that maintain the qualitative structure of temperature, differing in zero-point and scale.

More formally, 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, a is numerically greater than b.

3.13.1 Ordinal measurement

In this study, we assume that a single instance of a rating scale instrument results in *ordinal scores*.*Ordinality* implies that we know how elements are scored relatively to other elements. By meeting the axioms for ordinality, we say nothing about the ability of a score to adhere to . 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. Concatenation, or addition, requires additional assumptions. 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. A set of all tied elements with the same value is called an equivalence set. Let X, Y & X be any quantities in Q. The result of the projection of the quantity to an

ordinal scale, which we will refer to as P, needs to match 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);
- For strongly tied order: 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. Because the likert scale obviously allows for ties (multiple people can be ranked on the same score), and is thus a weak ordering, we assume transitivity and connectedness for the mapping. Transitivity implies that all order-relations need to be consistent. Connectedness implies that a connection between two elements is either larger, smaller, or equal. If antisymmetry is met, equal scores are only possible if they refer to the same object. In this case, the equivalence set has only one element.

3.13.2 Quantities and ratio-level measurement

As for the 'perfect' ratio-mapping of the principially measurable attribute we will assume that the following additional characteristics will hold above axiom 1 and 2 presented above (**?**).

- $X \oplus (Y \oplus Z) = (X \oplus Y) \oplus Z$ (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).

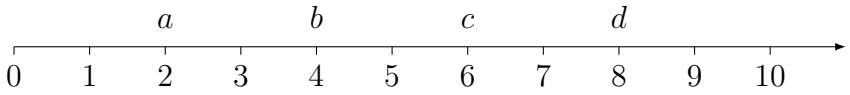
$\oplus$ refers to concatenation. This is the qualitative . Another consequences is that a value q can always be put in terms of another value r in Q. Every ratio-scale is homomorphic to the qualitative attribute that is being measured (**?**). The last axiom (the archimedean condition) is added to ensure 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 $\frac{x}{H}$ holds[3].

DATA AVAILABILITY STATEMENT

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

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.

[3] 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?


a
b
c
d
0
1
2
3
4
5
6
7
8
9
10