# Scales of Measurement

Nathaniel E. Helwig

University of Minnesota

## 1 What is Measurement?

What it means to "measure" something has long been a topic of both scientific and philosophical debate. The concept of measurement is fundamental to the field of psychology because we need reliable measurements of psychological constructs in order to trust any statistical results pertaining to those constructs. Despite the importance of measurement, this topic is often glossed over in many psychological applications—researchers often begin by *assuming* that they have measured their construct of interest, without necessarily providing any concrete evidence that such measurements are reliable or valid. Of course, this is a serious problem for interpreting results of psychological studies because statistical methods cannot overcome issues pertaining to poor measurement. More specifically, most statistical methods abide by the "garbage in, garbage out" principle, so you should expect to obtain invalid results if your input variables are measured inadequately.

In this chapter, we will not cover all of the specifics regarding psychological measurement—entire books and courses have been devoted to this topic. Instead, I will provide a brief overview of the "Theory of Scales of Measurement" that was proposed by Stevens (1946). In this influential paper, Stevens defined measurement as "the assignment of numerals to objects or events according to rules" (p. 677), and this broad definition still seems to be embraced by many applied psychological studies. In his paper, Stevens presents four different scales (or levels) of measurement that can characterize different types of measures that are used in psychological and other social science studies. It should be noted that Steven's approach to measurement has been widely criticized by researchers who specialize in measurement and statistics (e.g., see Michell, 1986). However, it is important to understand Steven's ideas, which are an implicit part of applied psychology.

# 2 Scales of Measurement

## 2.1 Nominal Scale

According to Stevens (1946), "[t]he *nominal scale* represents the most unrestricted assignment of numerals" such that "[t]he numerals are used only as labels or type numbers, and words or letters would serve as well" (p. 678). In other words, nominal scales of measurement involve assigning numerals that are *not* meant to convey any quantitative meaning. For example, suppose that we record the variable Gender, and code the responses as 1 = Female, 2 = Male, and 3 = Other. This would be an example of a nominal scale of measurement, given that the numbers 1, 2, and 3 are simply used as labels for the levels of Gender. In statistical language, variables that are measured using a nominal scale are discrete categorical variables that have probability mass functions.

## 2.2 Ordinal Scale

According to Stevens (1946), "[t]he *ordinal scale* arises from the operation of rank-ordering" such that "any 'order-preserving' transformation will leave the scale form invariant" (p. 679). In other words, ordinal scales of measurement involve assigning numerals that are only meant to convey meaning regarding the order of objects or events. Stevens correctly notes that "most of the scales used widely and effectively by psychologists are ordinal scales" (p. 679); however, psychological researchers typically treat them otherwise. As an example of an ordinal scale, think of the positions in which runners cross the finish line for a race, i.e., first place, second place, third place, etc. These positions can be used to put the runners in order, but they cannot be used for anything beyond ordering the runners. For example, using only the order that the runners crossed the finish line, we cannot say anything about differences in the runners's times or speeds—aside from the fact that the runner in position $i$ had a smaller time (or faster speed) than the runner in position $i + 1$. For variables with an ordinal scale of measurement, calculating difference scores, means, standard deviations, etcetera do not have any valid meaning. Instead, we should be focused on the quantiles of the distribution. In statistical language, variables that are measured using an ordinal scale are discrete (ordered) categorical variables that have probability mass functions.

## 2.3 Interval Scale

According to Stevens (1946), "[w]ith the *interval scale* we come to a form that is "quantitative" in the ordinary sense of the word. Almost all of the usual statistical measures are applicable here, unless they are the kinds that imply a knowledge of a 'true' zero point" (p. 679). In other words, interval scales are what we typically think of when we think of a quantitative measure, but such scales have a zero point that is "a matter of convention or convenience" (p. 679). The classic examples of interval scales of measurement are the Celsius and Fahrenheit scales that are used to measure temperature. Note that these scales have a linear relation to one another

$$^{\circ}\text{Fahrenheit} = ^{\circ}\text{Celsius}\left(\frac{9}{5}\right) + 32$$

and both scales have an arbitrary zero point. Regarding the arbitrary zero point, note that zero does not indicate a complete absence of the property being measured (i.e., temperature) for either scale: 0°C is when water freezes and 0°F is when a brine freezes.[1] Despite using ordinal measures used to collect psychological data, most psychological researchers treat their collected data as if were interval scale. In statistical language, variables that are measured using an interval scale are continuous variables that have probability density functions.

## 2.4 Ratio Scale

According to Stevens (1946), "*ratio scales* are those most commonly encountered in physics and are possible only when there exist operations for determining all four relations: equality, rank-order, equality of intervals, and equality of ratios" (p. 679). Note that ratio scales are similar to interval scales, except that ratio scales have a true zero point. As an example of a ratio scale, consider the measurement of the length of an object. In this case, we can convert between units by multiplying by a constant, for example

$$\text{foot} = 12\text{inch}$$

and such scales have a true zero point: 0 inches indicates a complete absence of the property being measured (i.e., length). Ratio scales are (almost?) never encountered in psychology.

[1] For some history of the Fahrenheit scale, see `https://en.wikipedia.org/wiki/Fahrenheit`

Table 1: Reproduction of Table 1 from Stevens (1946).

| Scale | Basic Empirical Operations | Mathematical Group Structure | Permissible Statistics (invariantive) |
|---|---|---|---|
| NOMINAL | Determination of equality | *Permutation group*<br>$x' = f(x)$<br>$f(x)$ means any one-to-one substitution | Number of cases<br>Mode<br>Contingency correlation |
| ORDINAL | Determination of greater or less | *Isotonic group*<br>$x' = f(x)$<br>$f(x)$ means any monotonic increasing function | Median<br>Percentiles |
| INTERVAL | Determination of equality of intervals or differences | *General linear group*<br>$x' = ax + b$ | Mean<br>Standard deviation<br>Rank-order correlation<br>Product-moment correlation |
| RATIO | Determination of equality of ratios | *Similarity group*<br>$x' = ax$ | Coefficient of variation |

*Note.* According to Stevens (1946) "any numeral, $x$, on a scale can be replaced by another numeral, $x'$, where $x'$ is the function of $x$ listed in this column" (p. 678).

## 2.5 Summary

Steven's scales of measurement are summarized in Table 1. The second column provides a simple description of what you can determine using each type of scale. The third column provides some details on how different units for a given scale must be related to one another. The fourth column describes what sort of analytic procedures are allowed for each scale of measurement. This fourth column is of particular importance for applications in psychology. As a reminder, most applied studies in psychology use ordinal measurements but apply methods that are only permissible for interval scale data. This is important because one could argue that the "reproducibility crisis" in psychology is related to the incongruence between psychological measurements and the methods applied to the measured data.

# 3 Reliability and Validity

## 3.1 Overview

Although not directly related to the scales of measurement proposed by Stevens (1946), the concepts of reliability and validity are essential to address whenever discussing the topic of measurement. In practice, it is important to understand the scale of measurement that is being used, as well as the quality of the measurements.

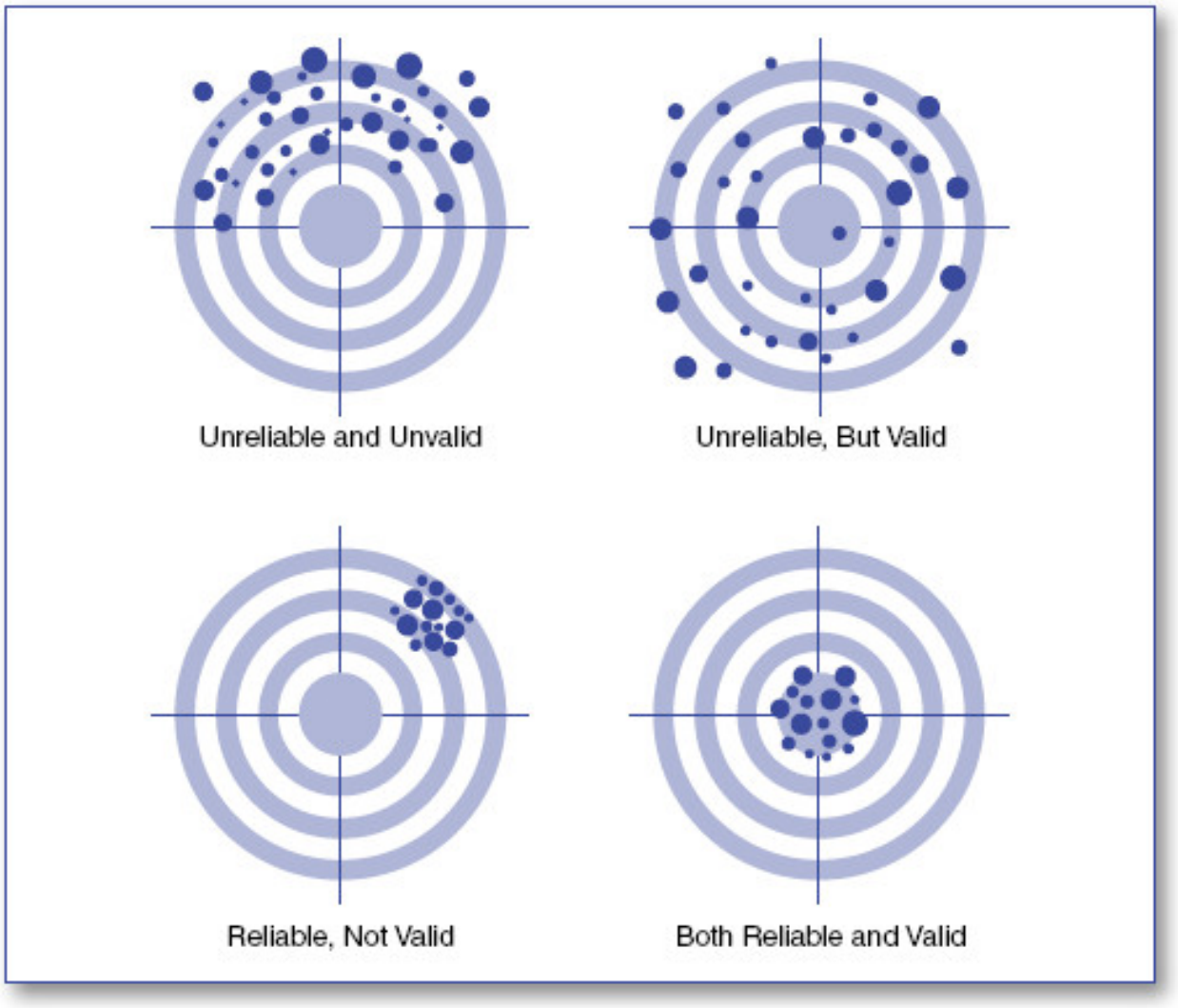


Figure 1: Visualization of reliability and validity from Ruel et al. (2016).

## 3.2 Reliability

There are many types of reliability that are discussed in psychological research, but all of the variants are centered around a common theme: a *reliable* measure is one that is "dependable, replicable, and consistent" (Ruel et al., 2016). In other words, a reliable measure is one that produces the same measurement results (up to the scale's precision) when measuring two objects that have the same amount of the property being measured. For example, if two individuals have the same weight, a reliable scale would return the same weight measurements up to the scale's measurement precision (e.g., 0.1 pounds).

Types of Measurement Reliability:

- *Test-retest reliability*: the correlation between two measurements of the same object measured at different times using the same scale.
- *Alternate form reliability*: the correlation between two measurements of the same object measured at the same time using different scales.
- *Internal consistency*: the pairwise correlations between the individual items that compose the measurement scale (item-wise congruence).
- *Split-test reliability*: the correlation between the scores on the first half and the second half of the measurement scale.
- *Inter-rater reliability*: the correlation between measurements as determined by two independent subjects (raters) measuring the same object.

## 3.3 Validity

There are many types of validity that are discussed in psychological research, but all of the variants are centered around a common theme: a *valid* measure is one that "operates the way [researchers] expect" (Ruel et al., 2016). In other words, a valid measure is one that measures what it is supposed to measure—without missing key properties or including unintended properties. For example, if an exam is supposed to measure statistical knowledge, then the exam would be a valid measurement if it comprehensively quantifies statistical knowledge without measuring extra unintended constructs (e.g., reading or language skills).

Types of Measurement Validity:

- *Face validity*: the measurement appears valid at face value.
- *Content validity*: the content of the measurement scale is complete, applicable, and representative of the measured construct.
- *Criterion-based validity*: the agreement between a scale's measurement and the measurement from a "gold standard" scale.
- *Concurrent validity*: the agreement between a scale's measurement and measurements of related (but distinct) constructs meaured from the same objects.
- *Predictive validity*: the ability of a measurement to predict related constructs.
- *Construct validity*: the degree to which a measurement scale is assessing the construct of interest, e.g., instead of some other construct.
- *Convergent validity*: the agreement between two measures in the same study that are intended to assess the same construct
- *Discriminant validity*: the lack of agreement between two measures in the same study that are intended to assess different constructs

# References

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin 100*(3), 398–407.

Ruel, E., W. Wagner III, and B. Gillespie (2016). The quality of measurement: reliability and validity. In E. Ruel, W. Wagner III, and B. Gillespie (Eds.), *The Practice of Survey Research: Theory and Applications*, pp. 78–100. SAGE.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science 103*(2684), 677–680.