# Scales of measurement and statistical analyses

Matt N. Williams
School of Psychology
Massey University
m.n.williams@massey.ac.nz

## Abstract

Most researchers and students in the social sciences learn of S. S. Stevens' classic taxonomy of four "scales" of measurement (nominal, ordinal, interval, and ratio). Many are exposed to the idea that conducting "parametric" analyses with ordinal or nominal variables is *inadmissible*, but are left confused about the actual basis for this rule (and why it is roundly ignored in practice). This is reflective of an unusual situation: Despite many researchers being *aware* of a rule proscribing particular forms of analysis with particular scales of measurement, the *basis* for this rule comes from a theory of measurement (representationalism) which is rarely discussed outside of the relatively esoteric body of literature concerned with measurement theory. Consequently, many struggle to achieve a satisfactory understanding of the basis for the rule. In this article, I attempt to provide an accessible and non-technical explanation of why the representational theory of measurement (arguably) implies that certain types of analyses are inappropriate with particular scales of measurement. I go on to describe alternative theories of measurement (operationalism, the classical theory of measurement, and latent variable theory), and suggest that latent variable theory more closely converges with how most contemporary researchers think about measurement than does representationalism. Furthermore, I explain how the measurement concerns that provoked Stevens' claims about inadmissibility are quite distinct from the *statistical* assumptions underlying data analyses—although there is an important connection between the two. I conclude by suggesting that researchers may find it more useful to critically examine their basis for assuming that particular attributes are *quantitative* and that particular relationships are *linear* rather than limiting themselves to the set of analyses that Stevens believed to be admissible with data of a given scale of measurement.

# Introduction

When students in the social sciences are introduced to research methods, once of the first concepts they are taught is often S. S. Stevens' (1946) division of measurements into four different scales of measurement: Nominal, ordinal, interval, and ratio. Stevens created this taxonomy in response to a long-running debate between psychologists and physical scientists concerning whether the data collection procedures taking place in psychology constituted *measurement* (see Ferguson et al., 1940). Stevens attempted to resolve this debate by suggesting that, instead of drawing a firm line between measurement and non-measurement, it could be possible to define measurement very broadly as "the assignment of numerals to objects or events according to rules" (Stevens, 1946, p. 677), but then divide measurement into four different "scales" (now more often referred to as "levels" of measurement). According to Stevens' definition and taxonomy of measurement, virtually any research discipline can claim to have achieved acts of measurement, although not all may achieve measurement at the interval or ratio levels. Stevens suggested, furthermore, that the scale of measurement with which particular variables have been measured then determines which statistical analyses are *admissible* with those variables.

Stevens' taxonomy went on to become extremely influential: His taxonomy is covered in many—perhaps even the vast majority—of research methods textbooks aimed at students in the social sciences (e.g., Cozby & Bates, 2015; Heiman, 2001; Judd, Smith, & Kidder, 1991; McBurney, 1994; Neuman, 2000; Price, 2012; Ray, 2000; Sullivan, 2001). The fame and influence of Stevens' taxonomy is something of a paradox in that his taxonomy forms part of an area of enquiry (measurement theory) with which—at least in my experience—few social scientists are acquainted. Measurement theories are theories directed at foundational questions about the nature of measurement. For example, what does it mean to "measure" something? What kinds of attributes can and can't be measured? And under what conditions can numbers be used to express relations amongst objects? Measurement theory can arguably be regarded as a branch of philosophy (see Eran, 2017), albeit one that often has heavily mathematical features.

Stevens' taxonomy is based on a specific theory of measurement (representationalism) that, while being possibly the most well-known theory of measurement, is nevertheless only occasionally discussed outside of the relatively esoteric measurement theory literature. It is also—as I will argue below—somewhat inconsistent with how most contemporary social scientists conceive of measurement. How has it come to be that a taxonomy based on a theory most have never heard of has become so influential? There are probably two major reasons. Firstly, Stevens' permissive definition of the term "measurement" is expeditious for social scientists, in that it sidesteps any debate about whether our disciplines can, in fact, claim to have achieved measurement (but see Michell, 2000, 2008). Secondly, Stevens' taxonomy is often used as the basis for heuristics indicating which statistical analyses should be used in particular scenarios (see for example Cozby & Bates, 2015).

Unfortunately, the fact that Stevens' levels of measurement are often described and taught without coverage of the related measurement-theoretic issues can mean that researchers are left confused about how to resolve common ambiguities. For example, researchers who collect data that seems obviously ordinal (e.g., responses to a rating scale) are often left confused about how it is that decision trees based on levels of measurement (e.g., Chatburn, 2017) suggest that only rank-based non-parametric tests are appropriate to use, while in actual practice almost all researchers apply parametric analyses with such data. Delving into the measurement theory literature sufficiently to satisfactorily resolve this confusion is no trivial task: The measurement theory literature contains several excellent resources pertaining to the topic of admissibility of statistical analyses (e.g., Gaito, 1980; Hand, 1996; Michell, 1986; Suppes & Zinnes, 1962), but this literature is often written for an audience of readers who are interested in measurement theory, and can be quite dense and challenging. Compounding the pedagogical problem further is the fact that more accessible resources directed at non- measurement theorists are of rather mixed quality. For example, two widely-cited articles by Carifio and Perla (2007, 2008) claim to resolve the issue of the admissibility of parametric analyses with rating scale data, but display almost little engagement with the extant measurement theory literature and confuse the *statistical* assumptions of parametric analyses with the *measurement theory* concerns that motivated Stevens' arguments about admissibility.

In this article, therefore, I attempt to provide an accessible introduction to the relationship between scales/levels of measurement and statistical analysis, with a focus on the question of which statistical analyses are appropriate with ordinal data (a common concern for psychologists).

## Statistical Assumptions

Before moving on to discussing measurement theories and their implications for statistical analyses, it is worth addressing the *statistical* (or “distributional”) assumptions of statistical analyses. Statistical assumptions are those assumptions formally drawn on when mathematically proving particular properties for particular statistics. For example, a statistical model commonly used by psychologists is the linear regression model, in which a participant’s score on a response variable is modelled as a function of his or her scores on a set of predictor variables multiplied by a vector of regression coefficients plus an error term $e_i$. In this model, the error terms $e_{1 \ldots N}$ are often assumed (over repeated sampling) to each be independently, identically and normally distributed with mean zero, regardless of the combination of levels of the predictor variables for each participant (Williams, Grajales, & Kurkiewicz, 2013). The assumption that the error terms have mean zero for any combination of values of the predictors is sometimes described as an assumption that the effects of the predictors are *additive* and *linear* (see Gelman & Hill, 2007).

By making particular assumptions, statisticians can prove mathematically (or demonstrate via simulation) that particular estimation methods have particular properties. For example, when the assumption above held, the estimation method of ordinary least squares provides regression coefficients that are unbiased, consistent, efficient, and normally distributed (see Williams et al., 2013 for definitions of these terms). Importantly, assumptions about scales or level of measurement are *not* usually invoked when statisticians develop statistical methods or demonstrate that statistical methods have particular properties.

Of course, the data collection methods that we would typically consider to produce ordinal data may very well sometimes lead to breaches of statistical assumptions (I will explain the term “ordinal” in more detail a littler later). For example, a Likert scale produces a discrete distribution of responses; thus, if a regression model estimated is estimated with a Likert scale as the response variable, the assumption of normally distributed errors cannot be met (since the normal distribution is a continuous probability distribution). This said, the effects of the breach of the assumption of normal errors for statistical inferences are likely to be limited to very slight changes to Type I and Type II error rates and confidence interval coverage (see Gelman & Hill, 2007), so this is typically a relatively minor concern outside of cases where the researcher is seeking to make predictions about individual cases.

Furthermore, a researcher who needs to deal with distributional problems but *isn't* concerned with whether her analyses are "admissible" according to SS Stevens' rules[1] has a large toolbox of methods available at her disposable to deal with those distributional problems. This toolbox includes (but is not limited to) bootstrapping and other resampling methods (Efron & Tibshirani, 1993; Good, 2013); Bayesian models in which parametric assumptions are deliberately loosened (e.g., by adding an estimated error skewness parameter to a model, rather than assuming that skewness = 0; see Arellano-Valle, Bolfarine, & Lachos, 2007); and other "robust" estimators (Wilcox, 2012);. In addition, distributional assumption breaches can be empirically detected (albeit always with some uncertainty attached), and their consequences empirically studied using simulations. As such, where distributional problems are the only concern, there is a good deal that researchers can do to address these problems, and for the most part they are probably best addressed on a study-by-study basis rather than via a general set of prohibitions. That said, there is at least one important way in which measurement concerns and statistical assumptions connect, and that is with respect to the common assumption of linearity in relationships; we will return to this assumption later.

## Representationalism and Scales of Measurement

Having covered statistical assumptions, and how they are distinct from measurement concerns, I will now turn to this article's main focus: measurement theory, and in particular the representational theory of measurement. The representational theory argues that measurement starts with a set of observable empirical relations amongst *objects*. The objects of measurement could literally be inanimate objects (e.g., rocks), but they could also be people (e.g., participants in a research study). In the theory, *measurement* consists of transferring the knowledge obtained about the empirical relations amongst objects (e.g., that granite is harder than sandstone) into a measurement scale which encodes the information obtained about these empirical relations (e.g., Mohs scale of mineral hardness). Representationalists do not necessarily assume that attributes (e.g., the hardness of rocks) have any independent existence; rather, objects exist, and relations amongst these objects may be observed. The representational perspective forms the basis of Stevens' (1946) four scales of measurement (nominal, ordinal, interval, and ratio).

[1] As we shall see shortly, the range of statistical analyses available to work with ordinal data, if following Stevens' rules about admissibility, is much smaller, being effectively restricted to rank-based non-parametric statistics.

**Nominal** measurement scales are produced when we have simply observed that some objects are not noticeably different, while other objects *are* noticeably different. For example, if we have a sample of participants, and we have noticed that some identify as men and some are women, and we code each woman as 1 and each man as 0 on our gender variable, then we have accomplished nominal measurement. Importantly, there are many such coding systems that would work just as effectively at conveying the information we have about participants' genders in this example. For example, we could just as well code women as 0 and men as 1, or women as -10 and men as 437.3745. As long as genders are recorded by assigning the men one fixed number and the women another, any two numbers would work just as well at conveying what we have observed about the participants. Formally, any coding system within the class of *one-to-one transformations* will equivalently convey the information that we have about the objects of measurement.

**Ordinal** measurement scales are produced when we have observed that some objects are noticeably greater than others with respect to some attribute.

In order to emphasise how the scale or level of measurement depends on the actual empirical observations obtained, I will use the same running example to convey ordinal, interval and ratio measurements. Imagine that Susan, a precocious 9 year old, has discovered in her parents' basement a set of old-fashioned beam balancing scales along with three balls: A rugby ball, a cricket ball, and a volley ball. Captivated by the scales, she sets out to compare the weights of the three balls. She quickly discovers that when she places the cricket ball in the cup on the left side of the beam, and the volley ball in the cup on the right side, the scale tips down to the right. Clearly, the volley ball is heavier. By a similar experiment, she discovers that the rugby ball is heavier than the volley ball. Unsurprisingly, she also discovers that the rugby ball is heavier than the cricket ball.

*Figure 1.* A set of balance scales. Photograph by user Poussin Jean [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0/)].

Susan now has some empirical observations about relations between objects that she could record using a measurement system. There are many possible coding systems that could convey the information she has collected, but there is a restriction: The number assigned to the rugby ball must be higher than the number assigned to the volley ball, which must be higher again than that assigned to the cricket ball. Any such system will record what Susan has observed: That the rugby ball is the heaviest of the three, followed by the volley ball, followed by the cricket ball. So, Susan could assign the rugby ball a weight of 3, the volley ball a weight of 2, and the cricket ball a score of 1. She could also assign the rugby ball a score of 1234, the volley ball 6, and the cricket ball 0.45. Either of these two systems[2] records the observed ordering *rugby ball* > *volley ball* > *cricket ball*. But she could not assign the rugby ball a score of 1, the volley ball a score of 2, and the cricket ball a score of 0: This coding system would not preserve the observed ordering, because it would imply that the volley ball is heavier than the rugby ball. Formally, any coding system within the class of *monotonic* transformations will equivalently convey the information that she has about the subjects (a monotonic transformation is one that preserves the original ordering of the observations).

**Interval** measurement scales are produced when, in addition to observing that some objects are greater than others with respect to some attribute, we have also been able to directly observe the ratios of the *differences* between objects (Suppes & Zinnes, 1962)[3].

Returning to Susan and her balance beam scales, imagine that Susan discovers a jar of 10-cent coins sitting near the balance beams. She realises that if she places the rugby ball on the left side of the beam, the volley ball on the right, and then adds coins into the bucket on the right, she can work out how many coins comprise the difference in weight between the volley ball and the rugby ball. This transpires to be 62 coins: With this number of extra coins on the right, the beam balances perfectly. She then sets out to compare the volley ball and the cricket ball, and discovers that she must add 31 coins to the side of the beam with the cricket ball in it in order for the weights to match.

Susan has now performed interval measurement: She is able to observe that the difference in weight between the volley ball and the rugby ball is twice[4] the difference in weight between the cricket ball and the volley ball. She can again use a numeric system to record what she knows about the weights of the balls, but there is now a new constraint:

[2] Like me, you may feel a vague sense of unease in relation to the latter coding scheme. I suspect that this unease stems from a sense that, even though the observations Susan has made thus far are solely about order relations, we know that the weights of the balls *in actual fact* stand in a particular set of ratios to one another, and we intuitively know that it is not plausibly the case that the rugby ball is 1234/0.45=2742 times as heavy as the cricket ball. The implicit measurement theory embodied in this perspective—the idea that measurement is not about encoding information about observed empirical relations, but rather about obtaining the best estimates of particular objects' values on underlying latent quantitative variables—is latent variable theory, which I will discuss later.

[3] In actual fact, it is also possible to produce interval measurement based only on observations about order and equality of differences along with some other conditions; see Suppes and Zinnes' (1962) description of infinite difference systems. For the sake of simplicity and brevity I have focused here on the simpler (albeit more restrictive) scenario of observations about ratios of differences. In actual practice, neither observations about ordering of differences nor ratios of differences are straightforward to produce for psychological attributes.

[4] As the example progresses, keen readers will realise that if the mass of each of the three balls were as specified by the relevant sporting governing bodies (155.9g for a cricket ball, 260-280g for a volley ball, and 460g for a rugby ball), then the ratios of differences Susan would find would not be quite as neat and round as they are presented here. We will attribute this to Susan's parents purchasing substandard sporting equipment.

Whatever the difference is between the number she assigns to the volley ball and the number she assigns to the rugby ball must be twice the difference between the numbers she assigns to the cricket ball and the volley ball. So she could code the weight of the cricket ball as 0, the volley ball as 1, and the rugby ball as 3; or she could code the cricket ball as 15, the volley ball as 30, and the rugby ball as 60. However, she could not code the cricket ball as 0, the volley ball as 1, and the rugby ball as 2: In this coding scheme, the observation that the difference between the volley ball and the rugby ball is twice that between the cricket ball and the volley ball would be lost. More formally, any coding system within the class of *linear transformations* will equivalently convey the information that Susan has about the weight of the balls.

**Ratio scales.** Imagine, now, Susan takes things one step further: She puts the cricket ball on the left side of the scale, and nothing at all in the right. She now starts adding coins until the beam balances. She discovers that the cricket ball is equal in weight to 47 coins. She then does the same for the volley ball (78 coins) and the rugby ball (140 coins[5]). She determines, thus, that the ratio of the weight the volley ball to that of the cricket ball is 1.66; similarly, the ratio of the weight of the rugby ball to that of the volley ball is 140/78 = 1.79, and the ratio of the weight of the rugby ball to the cricket ball is 140/47 = 2.98. Susan has now achieved ratio measurement: She is able to observe the ratios of the weights of the different balls to one another.

If Susan is now to code the weights of the balls and encode the information that she about the empirical relations between them, her choices are much more restricted. She can use the weight of the 10-cent coin as her unit of measurement and code the cricket ball as having a weight of 47, the volley ball as 78 (1.66 times the value of the cricket ball), and the rugby ball as 140 (2.98 times the value of the cricket ball). Or she could record the cricket ball as having a weight of 2, the volley ball as 3.32 (1.66 times 2), and the rugby ball as 5.96 (2.98 times 2). Both of these systems convey what is known about the weights of the objects: That the volley ball weights 1.66 times as much as the cricket ball is 1.66, and the rugby ball weights 1.79 times as much as the volley ball. But she could not code the cricket ball as 1, the volley ball as 2, and the rugby ball as 3; this coding system would not preserve the information she has collected about the ratios of the weight. More formally, only coding systems within the class of *multiplicative transformations* will equivalently convey the information that Susan has about the weight of the balls.

[5] Far be it from us to question how it is that Susan's parents possess such an excess of small change.

**Other scales of measurement.** Stevens' taxonomy is not exhaustive; there exist scales of measurement that do not fall neatly into his four categories, such as counted fractions (1993). The taxonomy is also fairly coarse; some of Stevens' categories can be separated into subtypes. For example, Stevens (1959, as cited in Eran, 2017) himself later divided interval scales into linear and logarithmic interval scales, and ratio scales into those with and without a natural unit ("absolute" scales; see also Suppes & Zinnes, 1962). A *count* variable (e.g., the number of suicides in New Zealand in one year) is an example of an absolute scale, where some specific number of events has been observed and represented in the form of a number.

## Representationalism and the Admissibility of Statistical Tests

The connection Stevens drew between his scales of measurement and statistical analysis was this. Given observations of a given level of measurement, there are a variety of coding systems that one could use to encode the information held about the empirical relations amongst objects. Furthermore, if we consider the four levels of measurement on a hierarchy from ratio at the top to nominal at the bottom, the lower levels of measurement offer a much more diverse range of coding schemes from which one can arbitrarily select. Furthermore, many statistical tests will produce different results depending on which coding scheme is used. If two different coding schemes produce differing statistical results, and yet the choice between the two coding schemes is purely arbitrary given the level of measurement, this is clearly not a satisfactory state of affairs, and implies that we should select a statistical analysis whose results do *not* depend on a purely arbitrary coding decision.

Stevens concluded that the only statistical analyses that are permissible for a sample of observations with a given measurement level are those that produce invariant results across the class of permissible (arbitrarily selected) transformations that can be used to record what has been observed about the empirical relations amongst objects. For example, imagine Susan has turned to considering her parents large collection of books, and now wants to determine whether the books in her Dad's enormous collection of historical fiction tend to be heavier than the books in her Mum's collection of sci-fi. Given that she is hoping to get back to the TV soon, Susan draws a sample of just 10 books from each collection (20 in total), and weighs pairs of books directly against one another only (i.e., returning to her ordinal approach to measurement, without any use of coins to determine ratios of weight or ratios of differences in weight). She is thus able to rank the weights of the 20 books in order, and can encode this information via any number of coding schemes: For example, she could code the lightest book as 1, the second-lightest as 2, and so on up in increments of one such that the heaviest book is coded as 20… or she could do almost exactly the same, but code the heaviest book as 47759, or she could use a third scheme where the two heaviest books are coded as having weights of 137 and 139, and so on.

Any of these schemes (or any other coding scheme within a monotonic transformation) would be just as effective at representing the information she has actually collected. However, it will be obvious that if Susan performs a Student's *t* test to compare the mean weights of the two samples of books, then the resulting *p* value is going to be rather different depending on which coding scheme is used. On the other hand, if she compares the two samples using a Mann-Whitney U test[6], the resulting *p* value will be identical regardless which of these coding schemes she uses, for the simple reason that the Mann-Whitney U is calculated using the *ranks* of the observations rather than their numerical coded values. As such, we might argue that the Mann-Whitney U statistic and its associated *p* value will be *invariant* across the class of permissible transformations in this example with ordinal data, whereas the Student's *t* test is not, and that as such the Mann-Whitney U is the most appropriate test.

Stevens went on to set out a list of statistical analyses that he believed to produce invariant results for variables of each scale of measurement. For example, he suggested that a median is appropriate as a measure of central tendency for an ordinal variable, since the case (or pair of cases) that falls at the median will always be the same across any monotonic transformation of the variable, even if the actual median itself will not[7]. On the other hand, he suggested that a mean is *not* a suitable measure of central tendency with ordinal data, because both the actual value of the mean and the case to which it most closely corresponds will both differ across different monotonic transformation of the data[8].

[6] Susan's primary school heavily emphasises statistical training, as all schools should.

[7] This line of reasoning holds well if the number of observations is odd; if the number of observations is even, then the median is usually defined as the mean of the two "middle" observations, and this median *will* depend on the coding scheme chosen to record the values.

[8] It is common for authors to claim that the issue of admissibility raised here implies that *parametric* statistical analyses should only be used with interval or ratio data (e.g., Jamieson, 2004; Kuzon, Urbanchek, & McCabe, 1996). In statistics, the terms "parametric" and "non-parametric" do not have unanimously agreed definitions (see Wasserman, 2006). However, broadly speaking a parametric analysis is one that involves an assumption that data or errors are drawn from a specific probability distribution (Altman & Bland, 2009). By this definition, some non-parametric tests (e.g., rank-based tests such as the Mann-Whitney U) produce invariant results across monotonic transformations of the outcome variable, and thus comply with Stevens' rules about admissible statistics. However, there certainly exist non-parametric tests that would not be considered as admissible by Stevens. For example, a permutation test to compare two means is non-parametric but will not

## Does Summing Ordinal Responses Produce an Interval Scale?

The argument has sometimes been made (e.g., Carifio & Perla, 2008) that while the responses to a rating scale item (e.g., a Likert item) are ordinal in nature, a score created by summing the responses to multiple items is itself interval in nature. This argument confuses the issue of levels of measurement with that of the *distribution* of a variable. It is true that the sum (or mean) of responses to items that individually have discrete ordered response options will usually have a distribution closer in shape to the normal distribution than that of the distribution of any of the individual items, due to the central limit theorem. However, this in no way implies that the sum-scores are interval, as no empirical observations have been made about the differences between objects or their ratios. It has been claimed in the past (Borgatta & Bohrnstedt, 1980) that having a normal distribution implies in of itself that a variable is interval, but this claim has been demonstrated to be false (Thomas, 1982): A variable can be normally distributed but not interval, and *vice versa*.

## Objections to Claims about Admissibility

A range of objections to Stevens' claims about the relationship between levels of measurement and admissible statistical analysis have been offered in the literature. A number of these criticisms are catalogued in Velleman and Wilkinson (1993). These criticisms include the fact that statistical analyses are not intrinsically based on assumptions about levels of measurement, as discussed above; the fact that there exist types of data that cannot be classified according to Stevens' categories; and the fact that treating data that is not strictly interval as ordinal means disregarding everything we know about the data points other than their ranked order, and thus disregarding potentially useful information. A particularly sophisticated discussion of the validity of Stevens' arguments can be found in Michell (1986), who points out that representationalism does not in of itself imply a need for a proscription of particular forms of analysis with data of particular levels measurement. Rather, the theory has implications about the conditions under which it is possible to use statements that are specific to a particular unit of measurement (e.g., "Gregor is 200cm tall and Tyrion is 100cm tall") to draw empirical conclusions that are *not* specific to a particular unit of measurement (e.g., "Gregor is twice as tall as Tyrion"). In this particular example, the

produce invariant results across monotonic transformations of the data. As such, Stevens' rules about admissibility are not accurately described as applying to whether or not "parametric" analyses can be utilised.

"scale-free" statement "Gregor is twice as tall as Tyrion" follows from the scale-specific version because height in centimetres is a ratio scale, but would not follow if we were measuring height using a nominal or ordinal scale. Setting aside specific criticisms of Stevens' claims about admissibility, though, one of the most fundamental issues is simply that the representational theory of measurement may itself not be representative of how researchers actually conceptualise of measurement, and what they hope to achieve by attempting to measure.

## Other Theories of Measurement

### Operationalism

Operationalism (see Bridgman, 1927; Chang, 2009) holds that an attribute is fully synonymous with the operations used to measure it: That if I say I have measured depression using score on the Beck Depression Inventory (BDI), then when I speak of a participant's level of depression I mean nothing more or less than the score the participant received on the BDI. As is the case for representationalism and Steven's scales of measurement, there is an echo of this measurement theory in the ubiquity of the idea of an "operational definition" as covered in almost any research methods textbook in psychology.

There are some fairly obvious problems with operationalism (see Borsboom, 2005). One problem is plurality of terms: If score on the BDI is depression, then a score on the Center for Epidemiologic Studies Depression Scale (CES-D) must measure some completely different construct from that measured by the BDI, which measures something different again from the Hamilton Rating Scale, and so on. Even a change in the items or response of the BDI would mean that we are dealing with an entirely different construct. Another problem is a conceptual framework for dealing measurement error, or the lack thereof: Because operationalism treats the construct and its operational definition as synonymous, it has no way of conceptualising measurement error. Measurement error cannot exist, because that would imply a difference between operational definitions and hypothetical constructs.

I suspect that few researchers would feel particularly comfortable about treating measurement error as a logical impossibility; as such, operationalism does not cohere very closely with how contemporary researchers typically think about measurement. Nevertheless, the operationalist perspective is useful to consider because it speaks to how the degree to which it is necessary to consider measurement issues when selecting statistical tests does depend on the nature of statistical inferences that are actually desired. For example, if a researchers hopes to make statistical inferences about the difference in mean BDI scores between men and women—and sees the issue of whether or not these inferences also hold with respect to *actual differences in depression level* between men and women as a matter of purely extra-statistical judgment—then there is little reason to be concerned about whether BDI scores are "ordinal" or "interval". I suspect that such a position would resonate with some researchers: For example, some would probably already assume that their statistical results will depend on the coding scheme they used to record data points, and not have any hope or expectation that their inferences would apply also to alternative coding schemes. In this sense, operationalism speaks to what sorts of statistical inferences we can (and cannot) justify if we are unwilling to consider measurement issues when selecting statistical tests.

## Classical Theory of Measurement

A third theory of measurement is what Michell (1986, 1993) terms the *classical* theory of measurement[9]. The term "classical" here refers to the fact that this theory of measurement has its roots in classical antiquity and the writings of Euclid and Aristotle. The classical theory of measurement can be stated as such: "When we measure in any department of natural science, we compare a given magnitude with some conventional unit of the same kind, and determine how many times the unit is contained in the magnitude" (Titchener, 1905, p. xix, as cited in Michell, 1986). For example, when Susan discovered that the cricket ball was equal in weight to 47 ten-cent coins, this was an act of measurement in exactly the classical sense: The weight of a ten-cent coin was the unit she used to express the weight of the cricket ball. It is only possible to completely express the magnitude of an attribute as a ratio of a common unit for some attributes, and we describe such attributes as *quantitative*. A quantitative attribute is *homogenous*: A weight of 10kg is not qualitatively different than a length of 1kg, but rather just more of exactly the same attribute.

A claim that an attribute is quantitative—at least in the sense the term "quantitative" is used in the classical theory of measurement—implies some very specific and falsifiable claims about the relations between units and objects. For example, if Susan discovers a gridiron football, and finds that it weighs the same as 126 coins, and she already knows that

[9] The classical theory of measurement should not be confused with classical test theory (see Lord & Novick, 1968), which is a theory not of the nature of measurement itself but rather of how variation in measurement outcomes can be attributed to different sources (true score and error).

the cricket ball weighs the same as 47 coins, then this implies that a combination of a cricket ball and 79 coins will weigh exactly the same as the gridiron ball. Susan could literally test this by putting both the cricket ball and 79 coins in one cup of her balance beam, and the gridiron football in the other cup; this act of literally adding up objects is known as *concatenation*. Barring operator error, she would no doubt find that the beam balances and the prediction is upheld.

Acts of concatenation are easy to achieve in relation to the simple physical attributes such as weight and length, but this is not the case for many other attributes, including psychological attributes. For example, imagine Joe has received a BDI score of 7 (minimal depression), Mikaere a score of 14 (mild depression), and Caitlin a score of 21 (moderate depression). If we take seriously the claim that the BDI *measures* depression level in the classical sense, this means that if we combined the depression levels of Joe and Mikaere we would end up with a person with exactly the same depression level as Caitlin. But we cannot concatenate people's levels of psychological attributes: There is no way to add Joe and Mikaere's actual levels of depression[10] together and then measure the level of depression of this combined person.

Indeed, when we consider the nature of a psychological attribute such as depression, it seems implausible to suggest that the difference between any two people's levels of depression can purely and completely be expressed in the form of a multiple of some constant unit. For example, the difference between Joe and Mikaere's depression scores might be explained by Mikaere having consistently low mood, while the difference between Caitlin and Mikaere's scores might be caused by Caitlin having begun to experience thoughts of suicide over the last fortnight—a qualitatively distinct symptom not experienced by Joe or Mikaere, and which may have qualitatively distinct causes from those that have precipitated Mikaere's low mood. If depression levels *cannot* be expressed purely as a multiple of some unit, then depression is not itself a *quantitative* variable, and consequently cannot be measured.

The German mathematician Otto Hölder lay out a series of axioms that define *quantity* in much more formal mathematical terms than I've presented here (see Michell & Ernst, 1996). Acts of concatenation are the simplest way to test whether these axioms hold

[10] We could add their *scores* on the BDI, but this isn't an act of concatenation: We would just be adding up the scores used to represent the participants' levels of depression, not their levels of actual depression. Compare this to the case of Susan literally combining the cricket ball and 79 coins in a single cup and then comparing the weight of this combination to that of the gridiron football.

for particular attributes, thus testing whether the attributes are indeed quantitative. However, a measurement model known as conjoint measurement makes it possible to test whether trios of attributes that are related to one another do in fact behave in accordance with the axioms of quantity (see Luce & Tukey, 1964). In other words, the hypothesis that a particular psychological attribute is quantitative is a testable[11] one, and Michell (2000, 2008) argues that psychologists should dedicate more time to testing whether psychological attributes are in fact quantitative (rather than just assuming that psychological attributes are quantitative and claiming to have measured them). Given that this article is intended to be a gentle introduction to measurement theories, the axioms of quantity and conjoint measurement theory are outside of its scope, but I encourage interested readers to refer to Michell (1999) for more reading.

## Latent Variable Theory

The final theory of measurement that I will discuss is latent variable theory (see Borsboom, 2008). In latent variable theory, the measurement outcomes (e.g., responses, scores, observations) are hypothesised to arise as the combined result of the effects of underlying *latent* variables along with *measurement error*. A participant's score on the BDI, then, might represent a combination of both some effect of his or her underlying level of *depression* (the unobserved latent variable about which we actually hope to make inferences), and also some effect of measurement error. Importantly, latent variable theory (like the classical theory of measurement, but unlike operationalism or representationalism) demands a *realist* ontological stance with respect to attributes: When we use latent variable models, we implicitly assume that the latent variables involved actually exist in the real world (Borsboom, Mellenbergh, & van Heerden, 2003).

[11] It's worth noting in passing that, while the conjoint measurement model permits a valid (albeit strict and challenging to implement) test of the hypothesis that a variable is quantitative, researchers have occasionally attempted other empirical tests of the hypothesis is quantitative (or interval). For example, a study of a "faces" rating scale measuring pain in children (Bieri, Reeve, Champion, Addicoat, & Ziegler, 1990) attempted to test whether children perceived there to be approximately equal differences in the pain displayed by the different faces in the response scale. A similar line of argument is used to claim that past empirical studies have demonstrated that Likert data can be interval in Carifio and Perla (2007). These attempts do not represent valid tests of the hypothesis that attributes are quantitative or interval. An interval scale is distinguished from an ordinal one by the type of observations that are represented in the measurements (observations about ordering vs. observations about ratios of differences), *not* by whether or not participants subjectively perceive there to be roughly equal distances between different points on the measurement scale itself.

A latent variable theory of measurement is implicit in popular statistical methods such as exploratory factor analysis, structural equation modelling (see Kline, 2015), and item response theory (see Embretson & Reise, 2000). These methods allow researchers to directly model relationships between latent variables and observed measurement outcomes. To my mind, latent variable theory is also the theory of measurement that is most consistent with how contemporary researchers actually tend to think about measurement: We assume that our observed measurement outcomes bear some relation to the variables we actually want to make inferences about (e.g., depression level), but we do not assume that this relation is perfect. The incredible popularity of structural equation modelling in contemporary research speaks to the appeal of latent variable theory; a Google Scholar search currently produces well over three million hits for the search string *structural equation modeling.*

From the perspective of latent variable theory, the fact that our measurement outcomes are produced by a combination of the effects of the latent variables we hope to make inferences about and measurement error implies that the statistical analyses we use need to account for the effects of measurement error. This is precisely what many applications of latent variables models are designed to achieve. This being said, the degree to which the results of a given analysis will be affected by measurement error will depend greatly on the type of statistical analyses, the research questions asked, the quantity and form of measurement error, and other factors. For example, Zumbo and Zimmerman (1993) show that a simple *t* test retains approximately its nominal Type I and Type II error rates even when comparing observed scores that are calculated by converting a latent variable to ranks (thus retaining only the observations about ordering) and adding measurement error. On the other hand, Westfall and Yarkoni (2016) show how substantially biased the coefficients of a multiple regression model can be when the predictor variables are measured with error. As such, whether it is necessary to explicitly model the effects of measurement error when taking a latent variable theory perspective is a contingent and semi-empirical question rather than one where it is possible to set firm and general rules about which statistics are admissible and inadmissible across a range of situations.

From a latent variable theory perspective, there is no direct need to classify variables according to their measurement level (nominal, ordinal, interval, ratio), or select analyses on this basis. However, the use of latent variable analyses with data that a representationalist might consider to be ordinal or nominal can result in at least two problems that are of more direct concern to a researcher taking a latent variable theory perspective. These two problems are, in fact, general measurement concerns that apply across many (but not all) statistical analyses. This is where will return, then, to statistical assumptions and their connection to measurement concerns.

## Measurement Theory Meets Statistical Assumption: Linearity and Quantity

Earlier in this article I pointed out that the statistical assumptions of particular statistical analyses are distinct from the measurement-theoretic concerns that inspired Stevens' arguments about the admissibility of statistical tests with data of differing levels of measurement. There is, however, an important way in which measurement-theoretic concerns have a direct bearing on the validity of the assumptions of many statistical analyses, including many latent variable models. This connection relates to the idea of a *quantitative* variable, as discussed in the section on the classical theory of measurement. Specifically, many (not all) commonly used statistical models assume relationships between variables that are *linear*. For example, in a simple linear regression model, the effect of a one-unit increase in the predictor is assumed to always result in the same change in the expected value of the response variable, regardless of where on the predictor scale that increase of one unit occurs. Similarly, in a confirmatory factor analysis model (a type of latent variable model[12]), we often assume that the relationship between each latent variable and each indicator is linear in nature[13]. But this assumption is only plausible if the predictor variables involved are quantitative[14]. If a predictor is ordered but not quantitative (i.e., a heterogeneous order; see Michell, 2012), then increases of the predictor are not increases of the same homogenous attribute but rather qualitatively different in nature from another. For example, we may think of depression as existing on a spectrum, but the symptoms that typically differentiate a person with mild depression from one without depression (e.g., persistent low mood, loss of pleasure) are often qualitatively different to the symptoms that will differentiate a person with severe depression from one without depression (e.g., suicidal ideation, psychomotor retardation, psychosis). If differences on a particular predictor variable are actually

[12] An interesting connection between measurement theory and latent variable modelling exists in the domain of item response theory (IRT) analyses. Specifically, some users of IRT produces estimates of participants' values on underlying latent traits, or conversion tables that allow raw scores to be converted to such estimates. Such conversions are sometimes thought to allow ordinal scores to be converted to interval ones (e.g., Harwell & Gatti, 2001). However, such a conversion is based on the *assumption* that the latent trait is itself quantitative, and this assumption is rarely directly tested. In the absence of evidence that the latent trait is quantitative, the claim that an IRT model can be used to convert ordinal data to interval is unjustifiable.

[13] We can admittedly loosen this assumption by treating observed variables as ordered categories; see Finney and DiStefano (2006). It is also even possible to apply latent variable models where the latent variables are ordinal classes rather than continuous or quantitative (see Jackson, Albert, Zhang, & Morton, 2013).

[14] Or dichotomous; the effect of a dichotomous variable on any other variable can never be anything but linear.

qualitative different sorts of things to one another, then it is hardly reasonable to expect changes in that predictor variable to have constant linear effects on an outcome variable. If we assume that a particular predictor has a linear effect on another, we are implicitly assuming that the predictor is quantitative (unless it is dichotomous, in which case it could only possibly have a linear effect).

We can extend this argument further beyond just linear relationships: Even where statistical analyses do not assume linear relationships between variables, they often assume that the relationships take some simple mathematical form (e.g., loglinear, quadratic, exponential). It is much rarer for researchers to estimate models where the relationship between one variable and another takes a non-smooth piecewise form, such that the effect on the outcome variable of an increment from 1 to 2 on the predictor variable could be completely different in magnitude and direction from the effect of a change from 2 to 3, and so forth. But this is exactly what we might expect if a predictor variable is a heterogeneous order—an “ordinal variable”—rather than a quantitative variable. See Figure 2.

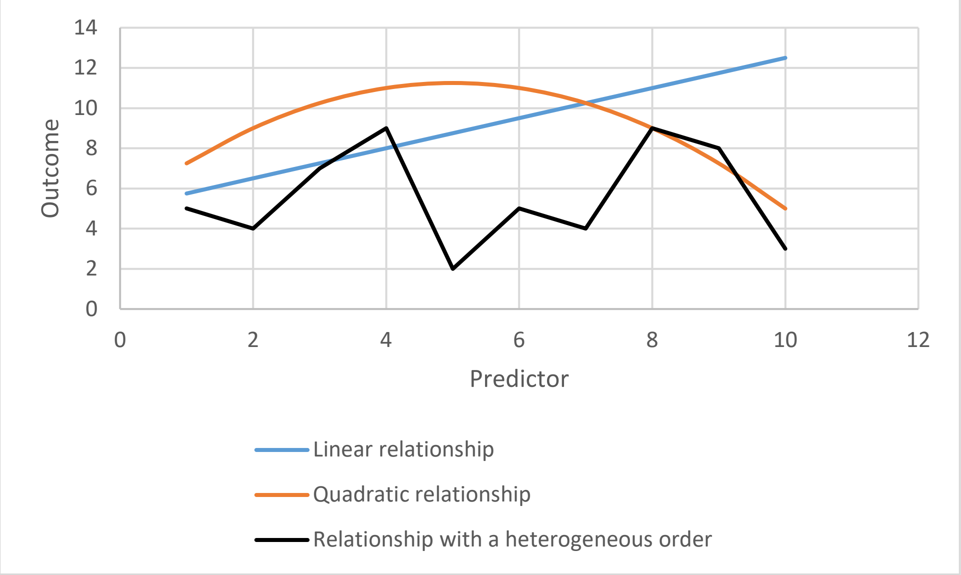

*Figure 2.* Illustration of three types of relationship: A linear relationship; a quadratic relationship (a relationship that is not linear, but still smooth), and finally the sort of relationship that is plausible for heterogeneous orders (where the effect of each increment can be completely different in direction and magnitude from the effects of other increments).

## Recommendations

At this point it should be clear that I see little strong reason for contemporary researchers to slavishly follow Stevens’ dictums about which statistical analyses are admissible with data of particular scales of measurement. Stevens' claims do not unambiguously follow from the measurement theory he based them on (representationalism). Furthermore, the implicit measurement theory of most contemporary researchers (latent variable theory) does not imply a need for the choice of statistical analysis to be based on the level of measurement of the variables. As such, I do not believe that it would be productive for contemporary researchers to follow Stevens' rules, with the ensuing restrictiveness that would thus apply to the analyses possible with most psychological data. However, it is clear that statistical assumptions and measurement-theoretic concerns do intersect in important ways. I therefore propose that researchers stop asking themselves "What level of measurement is my data, and how am I allowed to analyse it?", and instead ask these two questions:

The first question I suggest asking is *have I assumed that any relationships take a linear form, or some other simple mathematical form?* If so, what empirical or rational basis do I have for that assumption? (Recognising that an absence of evidence against non-linearity is not evidence *of* linearity).

If an assumption of linearity (or some other simple form of relationship) is not thought to be plausible as an actual description of literal reality, but rather forms a simplifying assumption for the sake of tractability, this should be clearly signalled to the reader. In cases where it is feasible to *avoid* assuming a simple mathematical form for a relationship—for example, by treating an indicator of a latent variable as categorical (Finney & DiStefano, 2006), or by applying ordinal-data models (Bürkner & Vuorre, 2019; Liddell & Kruschke, 2018)—these options should be seriously considered.

The second question I suggest asking is *have I implicitly assumed that any particular attributes or other variables are quantitative*? (If the answer to the first question is "yes", usually the answer to this one will also be yes). If I have made such an assumption, what *evidence* do I have to justify that assumption? Evidence for a claim that a particular attribute is quantitative may come from several fronts. For example, researchers in psychology may sometimes analyse physical variables that have been demonstrated to be quantitative by previous research in the natural sciences. Some of my own research (Williams, Hill, & Spicer, 2015a, 2015c, 2015b) has made use of temperature as a predictor; temperature is a physical variable and known to be quantitative (see Sherry, 2011 for a history). In other cases, it may be feasible to test the hypothesis that an attribute is quantitative using the conjoint measurement model (see Kyngdon, 2008, for an applied example). In other cases, the most that might be feasible is to *acknowledge* that a particular attribute (or set of attributes) has been assumed to be quantitative without evidence to support this assumption, and to recognise this as a limitation of the study.

## Conclusion

Contemporary researchers should not feel any compulsion to follow Stevens' rules about admissible statistics, but this does not free us from measurement concerns: The lack of evidence for the hypothesis that psychological attributes are quantitative has serious implications for statistical practice. I hope that this article will inspire more researchers to engage with the measurement theory literature, and to use it to inform their research practices.

## References

Altman, D. G., & Bland, J. M. (2009). Parametric v non-parametric methods for data analysis. *BMJ*, *338*, a3167. https://doi.org/10.1136/bmj.a3167

Arellano-Valle, R. B., Bolfarine, H., & Lachos, V. H. (2007). Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, *34*(6), 663–682. https://doi.org/10.1080/02664760701236905

Bieri, D., Reeve, R. A., Champion, G. D., Addicoat, L., & Ziegler, J. B. (1990). The faces pain scale for the self-assessment of the severity of pain experienced by children: Development, initial validation, and preliminary investigation for ratio scale properties. *Pain*, *41*(2), 139–150. https://doi.org/10.1016/0304-3959(90)90018-9

Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods & Research*, *9*(2), 147–160. https://doi.org/10.1177/004912418000900202

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge, UK: Cambridge University Press.

Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research & Perspective*, *6*(1–2), 25–53. https://doi.org/10.1080/15366360802035497

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Bridgman, P. W. (1927). *The logic of modern physics*. New York, NY: Macmillan.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. *Medical Education*, *42*(12), 1150–1152.

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116.

Chang, H. (2009). Operationalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2009/entries/operationalism/

Chatburn, R. L. (2017). Basics of study design: Practical considerations. *Cleveland Clinic Journal of Medicine*, *84*(Supp 2), e10–e19.

Cozby, P. C., & Bates, S. C. (2015). *Methods in behavioral research* (12th ed.). New York, NY: McGraw-Hill.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall/CRC.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New York: Lawrence Erlbaum Associates.

Eran, T. (2017). Measurement in science. In E. N Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2017/entries/measurement-science

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., … Tucker, W. S. (1940). Final report of the committee appointed to consider and report upon the possibility of quantitative estimates of sensory events. *Report of the British Association for the Advancement of Science*, *2*, 331–349.

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural Equation Modeling: A Second Course* (pp. 269–314). Greenwich, CT: IAP.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*(3), 564–567. https://doi.org/10.1037/0033-2909.87.3.564

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, United Kingdom: Cambridge University Press.

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses*. New York, NY: Springer.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *159*(3), 445–492. https://doi.org/10.2307/2983326

Harwell, M. R., & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, *71*(1), 105–131. https://doi.org/10.3102/00346543071001105

Heiman, G. W. (2001). *Understanding research methods and statistics: An integrated introduction for psychology* (2nd ed.). Boston, MA: Houghton Mifflin.

Jackson, J. C., Albert, P. S., Zhang, Z., & Morton, B. S. (2013). Ordinal latent variable models and their application in the study of newly licensed teenage drivers. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *62*(3), 435–450. https://doi.org/10.1111/j.1467-9876.2012.01065.x

Jamieson, S. (2004). Likert scales: how to (ab) use them. *Medical Education*, *38*(12), 1217–1218.

Judd, C. M., Smith, E. R., & Kidder, L. H. (1991). *Research methods in social relations* (6th ed.). Fort Worth, TX: Holt Rinehart and Winston.

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.

Kuzon, W., Urbanchek, M., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery*, *37*, 265–272. https://doi.org/10.1097/00000637-199609000-00006

Kyngdon, A. (2008). Treating the pathology of psychometrics: An example from the comprehension of continuous prose text. *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 108–113. https://doi.org/10.1080/15366360802035570

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*(1), 1–27. https://doi.org/10.1016/0022-2496(64)90015-X

McBurney, D. H. (1994). *Research methods* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*(3), 398–407. https://doi.org/10.1037/0033-2909.100.3.398

Michell, J. (1993). The origins of the representational theory of measurement: Helmholtz, Hölder, and Russell. *Studies In History and Philosophy of Science Part A*, *24*(2), 185–206. https://doi.org/10.1016/0039-3681(93)90045-L

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept*. Cambridge, United Kingdom: Cambridge University Press.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, *10*(5), 639–667. https://doi.org/10.1177/0959354300105004

Michell, J. (2008). Is psychometrics pathological science? *Measurement*, *6*(1–2), 7–24. https://doi.org/10.1080/15366360802035489

Michell, J. (2012). Alfred Binet and the concept of heterogeneous orders. *Frontiers in Quantitative Psychology and Measurement*, *3:*, 261. https://doi.org/10.3389/fpsyg.2012.00261

Michell, J., & Ernst, C. (1996). The axioms of quantity and the theory of measurement: Translated from part I of Otto Hölder's German text "Die axiome der quantität und die lehre vom mass." *Journal of Mathematical Psychology*, *40*(3), 235–252. https://doi.org/10.1006/jmps.1996.0023

Neuman, W. L. (2000). *Social research methods* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Price, P. (2012). *Research methods in psychology*. Washington, DC: Saylor Foundation.

Ray, W. J. (2000). *Methods: Toward a science of behavior and experience* (6th ed.). Belmont, CA: Wadsworth.

Sherry, D. (2011). Thermoscopes, thermometers, and the foundations of measurement. *Studies In History and Philosophy of Science Part A*, *42*(4), 509–524. https://doi.org/10.1016/j.shpsa.2011.07.001

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. https://doi.org/10.1126/science.103.2684.677

Sullivan, T. J. (2001). *Methods of social research*. Fort Worth, TX: Harcourt College Publishers.

Suppes, P., & Zinnes, J. L. (1962). *Basic measurement theory*. Stanford, CA: Stanford University.

Thomas, H. (1982). IQ, interval scales, and normal distributions. *Psychological Bulletin*, *91*(1), 198–202. https://doi.org/10.1037/0033-2909.91.1.198

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, *47*(1), 65–72. https://doi.org/10.2307/2684788

Wasserman, L. (2006). *All of nonparametric statistics*. New York, NY: Springer.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham, MA: Academic Press.

Williams, M. N., Grajales, C. A. G., & Kurkiewicz, D. (2013). Assumptions of multiple regression: Correcting two misconceptions. *Practical Assessment, Research & Evaluation*, *18*(11). Retrieved from http://www.pareonline.net/getvn.asp?v=18&n=11

Williams, M. N., Hill, S. R., & Spicer, J. (2015a). Do hotter temperatures increase the incidence of self-harm hospitalisations? *Psychology, Health & Medicine*, *21*(2), 226–235. https://doi.org/10.1080/13548506.2015.1028945

Williams, M. N., Hill, S. R., & Spicer, J. (2015b). The relationship between temperature and assault in New Zealand. *Climatic Change*, *132*(4), 559–573. https://doi.org/10.1007/s10584-015-1438-7

Williams, M. N., Hill, S. R., & Spicer, J. (2015c). Will climate change increase or decrease suicide rates? The differing effects of geographical, seasonal, and irregular variation in temperature on suicide incidence. *Climatic Change*, *130*(4), 519–528. https://doi.org/10.1007/s10584-015-1371-9

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology/Psychologie Canadienne*, *34*(4), 390–400. https://doi.org/10.1037/h0078865