Journal of Methods and Measurement in the Social Sciences,
Vol.10, No. 2, 76-94, 2019

Mathematization, Not Measurement: A Critique of Stevens' Scales of Measurement

M. A. Thomas[1]
Air Force Cyber College

In the early 1900s, physics was the archetypical science and measurement was equated with mathematization to real numbers. To enable the use of mathematics to draw empirical conclusions about psychological data, which was often ordinal, Stevens redefined measurement as "the assignment of numerals to objects and events according to a rule." He defined four scales of measurement (nominal, ordinal, interval, and ratio) and set out criteria for the permissible statistical tests to be used with each. Stevens' scales of measurement are still widely used in data analysis in the social sciences. They were revolutionary but flawed, leading to ongoing debate about the permissibility of the use of different statistical tests on different scales of data. Stevens implicitly assumed measurement involved mapping to real numbers. Rather than rely on Stevens' scales, researchers should demonstrate the mathematical properties of their data and map to analogous number sets, making claims regarding mathematization explicit, defending them with evidence, and using only those operations that are defined for that set.

Keywords: Measurement, Statistics, Social science methodology

Social scientists seeking to interpret data continue to rely heavily on a framework advanced by Stevens in 1946. Stevens (1946, p. 677) defined measurement as "the assignment of numerals to objects or events according to a rule." He set out four different scales of measurement (nominal, ordinal, interval, and ratio) and rules for determining the statistical tests that were permissible for each. Although Stevens' foundational paper is rarely taught, this paradigm remains the primary approach to measurement for the social sciences.

One of the leading critiques of Steven's framework is that his restrictions on the type of statistical tests to be used for different scales of measurement are unsupported. The academic debate has contributed to a varied practice. While some social scientists attempt to categorize their data as one of Stevens' four types and follow his guidelines regarding the permissible statistical tests to use, others disregard his limitations on permissible statistics.

Stevens' framework was revolutionary, yet flawed. Physicists and psychologists of the period equated measurement with mathematization to

[1] The author would like to thank Jeremy Martin for his patient accommodation of a total stranger who arrived without an appointment to discuss set theory. The views expressed are those of the author and do not necessarily represent views of the Department of Defense or its components.

real numbers. To expand the definition of "measurement," he invented a typology of scales and limits on "permissible" operations to help ensure that mathematical conclusions about data that lacked the mathematical properties of real numbers led to valid empirical conclusions. In so doing, he set the stage for continuing confusion among social scientists about the use of mathematical inference to draw conclusions about empirical data.

This paper describes Stevens' framework for measurement in the social sciences and the debate about the framework. It argues that Stevens' objective is better described as "mathematization" than "measurement." It explains how errors in mathematization can lead to incorrect empirical conclusions. It then traces some of the academic debate to the flaws in this early framework. The paper concludes by setting out the implications for quantitative social science research.

What Does It Mean to "Measure"?

In 1932 the British Association for the Advancement of Science appointed a committee composed of physicists and psychologists to evaluate whether "quantitative estimates of sensory events" were possible; or, in other words, whether such sensations were measurable. By "sensory events," they meant human sensations in response to physical stimuli, such as experiences of brightness when exposed to light or noise when exposed to sounds. On the meaning of "quantitative estimate" or "measurement," however, they were not able to come to any agreement. After eight years of debate, the committee concluded that "no practicable amount of discussion would enable them to express an agreed opinion" (Ferguson et al., 1940, p. 334).

Measurement was considered to be the bridge between the empirical world and the logic of mathematics. In the words of the physicist William Thomson (Lord Kelvin) in 1883:

> I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely in your thoughts advanced to the state of Science, whatever the matter may be. (1891, p. 80)

Psychologists believed that only if they could measure what they studied could they use mathematics to make inferences, and only if they could measure could psychology have the social status accorded to physicists, the archetypical scientists (Michell, 2005). However, human sensations could not be measured as the term had historically been defined.

"Measurement" meant the expression of the quantity of an attribute Y as a ratio of another quantity of that same attribute, X (see Michell, 2005; Ferguson, et al., 1940). The canonical example is the measurement of length. Lengths can be added together through the physical process of

laying rods end to end. The length of a rod *X* can be chosen as a standard to use to express the lengths of other rods and all other lengths can be expressed in terms of *X* as a ratio. To say that, for example, "the length of rod Y is two (in terms of *X*)" means that if two instances of rod *X* were laid end to end, their length together would be the same as that of rod *Y*.

Table 1
H. M. Johnson's Properties for Measurability

Additivity	The properties of two members can be physically added to be equivalent to the property of another member
Asymmetrical	The properties of two members, a and b, are such that either a > b, a = b, or a < b
Determinate	Measurability is limited by the ability to determine these equivalences (e.g., the accuracy of the balance)
Transitive	If a > b and b > c then a > c
Group Property	For the properties of any two objects, a and b, there is a third object for which the property is a + b; the number of objects that have the property is infinite
Commutative	If a + b = c then b + a = c
Associative	(a + b) + c = a + (b + c)
Axiom of Equals	If a = c and b = d then a + b = c + d and a + d = b + c
Neutral Member	If there is an opposite of the property, then there must be an object that has none of the property, z, such that z + a = a + z = a
Negative	For every object with property a, there must be an object with a property of negative effect, b, such that a + b = z (the neutral member)

Note. Adapted from H. M. Johnson (1936)

The physicist Campbell referred to this definition as "direct" measurement (and later, "fundamental" measurement). He distinguished it from "indirect" (later "derived") measurement," which he defined as attributes that are mathematically derived from fundamental measures, such as area (a function of length), or density (a function of mass and

volume) (Ferguson, et al. 1940; Campbell, 1957) Quantities were then mapped to a numeral sequence or "scale" in such a way that fixing the first point of the scale determined the rest.

This concept of measurement implies that to be measurable, a thing must empirically demonstrate certain mathematical properties. For Campbell, some physical process of addition or concatenation was required for an attribute to be measurable. H. M. Johnson (1936), a critic of early psychometrics, described in detail the mathematical properties that the subject of study must demonstrate empirically for a thing to be measurable. (Table 1).

This posed little problem for physicists, who worked primarily with attributes that had these properties, but most of the work of psychologists involved the discovery or exploration of relationships of order. For example, when exposed to noise, subjects could rank their sensations of sound in terms of greater or lesser, but it could not be demonstrated empirically that two noises of the same loudness created twice the sensation of sound. Human sensations were neither countable nor additive and could not be expressed as a ratio.

Early psychologists wrestled with questions regarding the nature of their data and whether and how that data might be measured (see, e.g., Thurstone, 1929; Johnson, 1936). They proposed to widen the definition of measurement. This proposal provoked vociferous objections, such as Guild's response, "In Denial of the Measurability of Sensation Intensity":

> There is no doubt that the desire of psychologists to be able to apply the processes and concepts of measurement to the field of sensory experience is due to the success of such processes in physics and geometry, and that their aim is to introduce the same kind of definitiveness into descriptions of sensory behaviour that quantitative laws give to descriptions of physical phenomena. . . . What some of them (not all) appear to be unable or unwilling to do is to realize that they cannot obtain this significance in relations involving numbers derived from processes of types differing fundamentally from those which form the basis of all physical measurements. To insist on calling these other processes measurement adds nothing to their actual significance but merely debases the coinage of verbal intercourse. Measurement is not a term with some mysterious inherent meaning, part of which may have been overlooked by physicists and may be in course of discovery by psychologists. (Ferguson et al., 1940, p. 345)

However, drawing on Campbell, Stevens' 1946 paper advanced a new definition of "measurement" and a way to measure psychological data. His framework became canonical in psychometrics, and from there in statistics for the social sciences more broadly. He defined measurement as "the assignment of numerals to objects and events according to rules" (Stevens, 1946, p. 677). Different rules of assignment lead to different "scales" or "levels" of measurement: nominal, ordinal, interval, and ratio. (Table 2). The problem then becomes that of making explicit (a) the various rules for the assignment of numerals, (b) the mathematical properties (or group

structure) of the resulting scales, and (c) the statistical operations applicable to measurements made with each type of scale (Stevens, 1946).

In Stevens' framework, researchers select the appropriate scale of measurement based on their ability to demonstrate the mathematical properties of the subject of study. A researcher with a means to determine

Table 2
Stevens' scales of measurement (Stevens 1946, p. 648)

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (Invariantive)
Nominal	Determination of equality	*Permutation group* x'=f(x) f(x) means any one-to-one substitution	Number of cases Mode Contingency correlation
Ordinal	Determination of greater or less	*Isotonic group* x'=f(x) f(x) means any monotonic increasing function	Median Percentiles
Interval	Determination of equality of intervals or differences	*General linear group* x'=ax+b	Mean Standard deviation Rank-order correlation Product-moment correlation
Ratio	Determination of equality of ratios	*Similarity group* x'=ax	Coefficient of variation

whether two quantities are equal can create a nominal scale; a researcher with the means to determine whether one quantity is greater than another can create an ordinal scale; a researcher with the means to determine if differences in quantity are equal can create an interval scale; and the researcher with the means to tell if two ratios of a quantity are equal can create a ratio scale.

Stevens also specified what he called "permissible statistics" for each scale. His criterion for permissible statistics was that the mathematical operations yield answers that were independent of the arbitrary representational choices of the researcher. For example, data measured on an ordinal scale should yield the same result for any order-preserving transformation of the scale; permissible operations include calculation of medians or percentiles.

Starting in the late 1950s, mathematical psychologists built on the work of Stevens and others to develop Representational Measurement Theory (RMT) (see, for example, Luce and Tukey, 1964; Luce and Suppes, 2002). RMT defines a "theory of measurement" as "a precise specification of how a scale is formed" where a "scale" is "a set of structure preserving mappings . . . from the qualitative or empirically based structure into a structure from pure mathematics" (Narens, 2002, p. 757).

Stevens' framework is still used and debated. The highly abstract and mathematical nature of the RMT literature makes it inaccessible to most. This is one reason that RMT has had little impact on practitioners even in its originating field of psychology (Boumans, 2016). Cliff (1992) wrote that "even quantitatively sophisticated areas of psychology behave as if abstract measurement theory did not exist" (p. 187). Michell (2008) wrote that measurement theory is "excluded from consideration in mainstream psychometrics" and "missing from the curriculum" (pp. 8-9). An evaluation of RMT is outside of the scope of this paper.

Permissible Statistics

Stevens' clam that only some statistical operations are permissible, depending on the scale, has led to continuing debate. Critics took issue with the idea that either Stevens or empirics could create constraints in mathematics. Gaito (1980) wrote:

> In mathematical statistics literature one will not find scale properties as a requirement of the use of the various statistical procedures. This requirement was merely a figment of the imagination of a number of psychologists because of a confusion of measurement theory and statistical theory. Statistical procedures do not require specific scale properties. The assumptions for the use of statistical procedures can be clearly stated and are based on the mathematical aspects underlying the procedure. (p. 566)

Anderson (1961, p. 309) argued that the *F* and *t*-tests can be applied to ordinal scale data, writing that "the validity of a statistical inference cannot depend on the type of measuring scale used." Burke (1953, p. 73) wrote that "the use of the sample mean and standard deviation does no violence upon the data, whatever the properties of the measurement scale. Thus, the use of the usual statistical tests is limited only by the well-known statistical restrictions." McRae (1988, p. 162) wrote, "Mathematical statistics is defined within the domain of numbers. Its operations and its results are confined to that domain. The validity of statistical results *as numbers* does not depend on any correspondence between the numbers and objects and events in the world." Even Michell (1986), who was more sympathetic and presented a different argument supporting Stevens' conclusions, called Stevens' assertions about permissible statistics "high handed."

The debate about permissible statistics has often focused on the proper characterization of Likert items and Likert scales, proposed by Rensis

Likert, which are ubiquitous in the social sciences and were originally proposed to measure attitudes and perceptions (Likert, 1932). Likert items are survey questions in which respondents are asked to choose an answer from (typically five or seven) ordered responses. Each possible answer is numbered. Likert scales are derived by summing or averaging the answers to similar Likert items.

Social scientists disagree about the scale of measurement of these data in theory and treat them differently in practice. Jamieson (2004) argued that Likert scales are ordinal data, yet are commonly treated as interval data. This matters because "if the wrong statistical technique is used, the researcher increases the chance of coming to the wrong conclusion about the significance (or otherwise) of his research" (Jamieson, 2004, p. 1217). Kero and Lee (2016) agreed, in their article titled "Likert is Pronounced 'LICK-urt' not 'LIE-kurt' and the Data are Ordinal not Interval." Carifio and Perla (2007, 2008) argued that while Likert items are ordinal, Likert scales are interval if not ratio data and that their interval nature is an empirical fact, an "emergent property" of the scale built from Likert items. They in turn cited Glass, Peckham, and Sanders (1972, p. 237), who argued that "[t]he relevant question is not whether [analysis of variance] assumptions are met exactly, but rather whether the plausible violations of the assumptions have serious consequences on the validity of probability statements based on the standard assumptions." Norman (2010) agreed with Jamieson that Likert scales are ordinal, saying that the matter "does not take a lot of thought." He nevertheless defends the calculation of "change scores" (measures of the difference in the value of a variable over time) for ordinal data, arguing,

> One of the beauties of statistical methods is that, although they often involve heroic assumptions about the data, it seems to matter very little even when those are violated. . . . If Jamieson and others are right and we cannot use parametric methods on Likert scale data, and we have to prove that our data are exactly normally distributed, then we can effectively trash about 75% of our research on educational, health status and quality of life assessment . . .. (p. 627)

Norman (2010) concludes that "[s]ince an ordinal distribution amounts to some kind of nonlinear relation between the number and the latent variable" then given that analysis of variance (ANOVA) is robust with respect to non-normality, it can be used with ordinal data (p. 627).

The academic debate about permissible statistics is mirrored in a highly varied statistical practice. Social scientists often ignore Stevens' equivocal limitation on permissible statistics, for example, by calculating the means of ordinal data. Likert scales are routinely calculated from Likert items and subjected to parametric statistical analysis. Even more frequently, social scientists do not specify the scale of data they are analyzing at all. Citing statistical textbooks, Jamieson cautions that the calculation of means and standard deviations are "inappropriate" for ordinal data:

> However, these rules are commonly ignored by authors, including some who have published in Medical Education. For example, the authors of 2 recent papers had used Likert scales but described their data using means and standard deviations and performed parametric analyses such as ANOVA. This is consistent with Blaikie's observation that it has become common practice to assume that Likert-type categories constitute interval-level measurement. Generally, it is not made clear by authors whether they are aware that some would regard this as illegitimate; no statement is made about an assumption of interval status for Likert data, and no argument made in support. All of which is very confusing for the novice in pedagogical research. What approach should one take when specialist texts say one thing, yet actual practice differs? (Jamieson, 2004, p. 1217)

The Problem with the Scales of Measurement

Physicists who held what Michell (1986) called the "classical" view of measurement saw measurement as the recording of empirical facts. By contrast, Stevens saw mathematics as analogy and "measurement" as mapping from empirical data to different types of number series based on the researcher's ability to demonstrate the empirical qualities of their data. The researcher could then perform mathematical operations on the numbers and use the conclusions to draw empirically valid conclusions about the subject of study.

> Scales are possible in the first place only because there is a certain isomorphism between what we can do with the aspects of objects and the properties of the numeral series. . . . The isomorphism between these properties of the numeral series and certain empirical operations which we perform with objects permits the use of the series as a model to represent aspects of the empirical world. (Stevens, 1946, p. 677)

Stevens had two objectives: to show that psychologists could measure their subjects of study and therefore that psychology was a science and to use mathematics to draw empirical conclusions. He attempted to expand the definition of "measurement," but as a practical matter, what he sought was "mathematization": "to reduce to mathematical form or subject to mathematical treatment" (Merriam-Webster, 2019). However, no distinction had yet been made between the two. Because physicists worked with properties that could be expressed in ratios, measurement was equated by both physicists and psychologists without discussion as mathematization to rational or real numbers.[2]

To prevent researchers from conducting mathematical operations defined on the real numbers for data that did not evidence the mathematical properties of real numbers, Stevens created different types of scales and

[2] Although the discussion focused on whether quantities of attributes could be added or expressed as ratios, in practice mathematical operations were often used that are not defined on rational numbers, such as taking the square root of a number without considering whether it is a perfect square.

"permissible statistics." He did not explicitly anchor his discussion in mathematics, instead presenting the limitations as fiat. He also equivocated, suggesting that these limitations were at the researcher's discretion. For example, he argued that while taking the mean and standard deviation of ordinal scales should not be done "in strictest propriety" nevertheless "[i]n numerous instances it leads to fruitful results" (Stevens 1946, p. 679). The degree of impropriety depended on the extent to which "successive intervals on the scale are unequal in size."

Researchers wrestled with the choice of using an ordinal scale, with information loss and a more limited ability to draw mathematical inferences, or using a ratio scale, creating a logical leap between the data and its mathematical representation and casting doubt on the validity of any empirical conclusions. As one statistics instructor blogged:

> All ordinal data are not the same. There is a continuum of "ordinality" if you like. . . . There are some instances of ordinal data which are pretty much nominal, with a little bit of order thrown in. . . .The mode is probably the only sensible summary value other than frequencies. . . . Then there are other instances of ordinal data for which it is reasonable to treat it as interval data and calculate the mean and median. It might even be supportable to use it in a correlation or regression. This should always be done with caution, and an awareness that the intervals are not equal. . . However, at the same time as saying, "you should never calculate the mean of ordinal data", it would be worthwhile to point out that it is done all the time! (Petty, 2013)

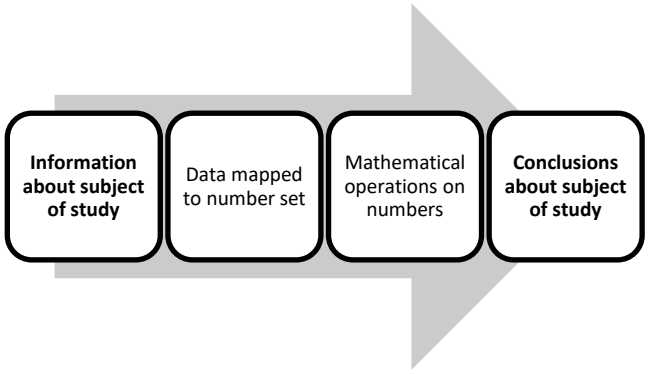
Both the creation of this fiat structure and Stevens' departures from it set the stage for continuing controversy. However, aside from political reasons, there was no need for Stevens to invent a structure to account for the differences between the mathematical properties of the data and those of the real numbers. Set theory was developed by Georg Cantor in the late 1800s, and gradually became the basis for modern mathematics. With it came the notion of different sets with different properties, such as unordered sets, ordinal numbers, integers, rational numbers, and real numbers.[3] Data can be mathematized to the appropriate set based on their empirically demonstrated mathematical properties, rather than to one of Stevens' "levels of measurement." In mathematics, operations are defined, not labeled as "permissible" by Stevens. If undefined, then they are mathematically meaningless. Whether mathematical conclusions in turn result in valid *empirical* conclusions depends both on the validity of the mapping operation and the soundness of the mathematical logic (Figure 1.)

In this light, the academic discussion about permissible statistics referenced earlier is both right and wrong. The validity of statistical inference cannot and does not depend on what type of scale is used, because a scale is not a mathematical object. However, statistical tests may only be defined for certain number sets. For example, ordinals do not have

[3] Readers interested in a mathematical introduction to naïve set theory are referred to Shen and Vereshchagin (2002).

intervals, and so it is mathematically meaningless to discuss the sizes of their intervals or to calculate their means. Moreover, errors of mathematization can ultimately undermine the validity of the final empirical conclusions.

Figure 1. Empirical information about the subject of study is mapped to an analogous number set; mathematics is used to draw conclusions about the numbers; mathematical conclusions are used to draw empirical conclusions about the subject of study. The validity of the final empirical conclusions depends on correct mathematization of the subject of study and sound mathematical reasoning.



Bad Mathematization

When data are mathematized to sets that have different or additional properties, and mathematical operations that depend on those properties are carried out on those numbers, the results are mathematically valid but do not translate to valid empirical conclusions. Three examples of this confusion are offered: Stevens' own inclusion of labeling as a form of measurement; Lord's famous critique of Steven's limits on permissible statistics, which focuses on the permissible mathematical operations for nominal numbers; and the current confusion about the treatment of Likert items.

Labeling

Consistent with his definition of measurement as "the assignment of numerals to objects or events according to a rule," Stevens including labeling data using numerals as names as a form of measurement, even though he acknowledged that some would find it "absurd." (Stevens 1946, p. 679) An example would be the assignment of numbers to highways or football players for ease of reference.

It is possible to map members of an unordered set ("A") to members of a set of numbers ("B"). However, this would not lead to useful empirical

conclusions about A. If B is also an unordered set, for example, a set of numerals that are simply graphic symbols, the same operations may be conducted equally easily on A as on B and there is no reason to manipulate B instead of manipulating A directly. It is just as easy to count the total number of highways as to count the total number of highway numerals. Steven's permissible statistics for the nominal scale can be performed directly on A. Moreover, the use of a set of numerals as names or graphic symbols rather than as numbers may lead to honest errors by people who think they are intended to represent a number set.

If B is some other type of number set, with mathematical properties that A does not have, mathematizing A to B is an error and the performance of mathematical operations on B that are not defined for A would not lead to valid empirical conclusions. For example, one could mathematize the set of household pets to real numbers, mapping the cat to "10" and the dog to "2." It is possible to take the mean of the real numbers 10 and 2, and come to a conclusion about B, the set of numbers to which A is mapped, such as that the average of B is six. However, this mathematical conclusion does not lead to an empirical conclusion about A, the set of pets. A dog and a cat cannot be averaged, and the statement that the average of the household pets is six is a meaningless and arbitrary artifact of the researcher's mapping choice; the researcher could as easily have named them "10" and "12." To prevent this, calculation of means is not permitted for the nominal scale under Stevens' framework.

Football Numbers

One of the most well-known critiques of Stevens' scales is a highly cited 1953 paper, "On the Statistical Treatment of Football Numbers," by the psychometrician Frederic Lord (1953). The critique is flawed. Lord's argument illustrates one or both of the above errors: treating numeric labels as numbers, or mathematizing an unordered group to a number set with different properties.

Lord argued that, contrary to Stevens' limits on permissible statistics, all statistical operations are permitted even on nominal numbers. To make his point, he tells a parable about a machine that dispenses numbers without replacement that are used to identify football players. The freshmen complain that their numbers are too low compared to the numbers given to the sophomores and that someone has tampered with the machine. The hero of the parable demonstrates that there is nothing wrong with the machine by calculating the critical ratio to evaluate the probability that the sample of freshman football numbers was randomly drawn. To calculate the critical ratio, the hero adds, multiplies, and even divides nominal numbers over the horrified gasp of the professor. Lord claims that the analysis demonstrates that the sum of a sample of football numbers "obeys the same laws of sampling as they would if they were real honest-to-God cardinal

numbers" because "the numbers don't remember where they come from" (Lord, 1953, p. 751).

The numbers may not know where they come from, but the researcher and statistician must know their properties. Lord does not specify what type of numbers are dispensed by the machine, but he performs operations on them as if they were randomly drawn, normally distributed real numbers. If the football numbers are of some other type—for example, if they are integers, for which division is defined differently, or numeric labels, not numbers—then the mathematical operations his hero attempts to perform are undefined in mathematics and so do not produce the results listed in the paper.

Even if they are real numbers for which such operations are defined, the mathematical conclusions are about the numbers, not the football players. This is not evidence that one can perform such operations on an unordered set (that of football players); division of unordered sets is undefined in mathematics. Lord was correct that one can calculate critical ratios using nominal data, only if by "nominal data" he meant normally distributed, randomly drawn real numbers to which an unordered set has been mapped. Yet, most researchers would consider the "nominal data" to be the unordered set itself.

A similar problem is presented when relationships of order are mathematized to real numbers and operations that are only appropriate for real numbers are performed on them. For example, imagine that there are two groups of people, Group A and Group B. Each group has five people in it with different weights, and they are assigned ordinal numbers according to weight, with 1 being the lightest and 5 being the heaviest. If the researcher then maps the ordinal numbers to real numbers and calculates means, the average of A and B is the same, three. However, this is not an average group weight, but an average of the real numbers to which the order numbers were mapped. It does not mean that the average weight of the two groups is the same. It does not provide any information about the weights of the two groups or translate into any empirical conclusions about group weight. Moreover, there are an infinite number of ways that these ordinals could be mapped to real numbers at researcher discretion as long as the order is preserved (for example, 1 is mapped to 100, 2 is mapped to 139, etc.), making the calculated mean a function of this arbitrary choice by the researcher.

Likert Items and Scales

Although most researchers do not need to be convinced that they should not attempt to calculate the averages of highways, cats, dogs, or football

players, the same issues arise in the discussion about the treatment of Likert items and Likert scales.

Multiple-choice questions can be used to collect data with different properties, including data that can be mathematized to unordered, ordinal, cardinal, or real number sets.[4] Some members of the set of possible responses are selected and presented to the respondent to allow the respondent to choose. The respondent's choices are hoped to provide information about the underlying subject of study. The respondent's possible answers are labeled with numbers: the numbers one through five in the case of a five-option question. In some cases, mathematical operations are then performed on the answer numbers in order to draw empirical conclusions about the subject of study (Figure 2). Unfortunately, researchers rarely make explicit claims about the mathematical properties of their subject of study, provide evidence of those claims, or describe their mathematization of the answer numbers.

For mathematical operations on the answer numbers to lead to conclusions about the subject of study that are *empirically* valid, the operations must depend only on mathematical properties that are evidenced by the original data. Because multiple-choice questions may be used to gather data about subjects of study with different mathematical properties, the appropriate mathematization of the answer numbers will vary with the empirically demonstrated mathematical properties of the subject of study. For example, if the data are members of an unordered set, treating the answer numbers as numbers is problematic because they are ordered.

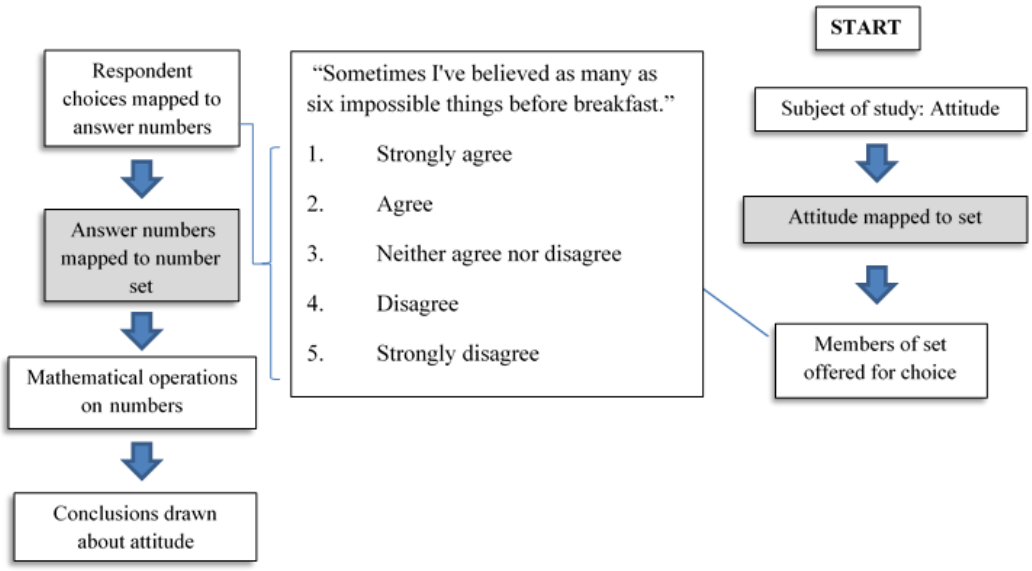
The first problem that gives rise to the confusion about Likert items and scales, originally designed to measure attitudes and perceptions, is that the mathematical properties of attitudes and perceptions were assumed in order to facilitate the use of mathematics, not empirically demonstrated. Thurstone and Chave (1929) made an early case for the mathematization of attitudes. Like their contemporaries, they equated measurement with mathematization to real numbers and argued that attitudes could be measured in this way:

> When we discuss opinions, about prohibition for example, we quickly find that these opinions are multidimensional, that they cannot all be represented in a linear continuum. The various opinions cannot be completely described merely as "more" or "less." They scatter in many dimensions, but the very idea of measurement implies a linear continuum of some sort such as length, price, volume, weight, age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind. We judge in a similar way qualities such

[4] It is possible to use multiple-choice survey questions to gather data that is best mathematized as cardinal or real numbers, but in practice it is rare. It is easier to simply ask the respondent for the number rather than to tee up limited choices from which the respondent must select, and which may then be an approximation of the correct answer.

> as mechanical skill, the excellence of handwriting, and the amount of a man's education, as though these traits were strung out along a single scale, although they are, of course, in reality scattered in many dimensions. As a matter of fact, we get along quite well with the concept of a linear scale in describing traits even so qualitative as education, social and economic status, or beauty. A scale or linear continuum is implied when we say that a man has more education than another, or that a woman is more beautiful than another, even though, if pressed, we admit that perhaps the pair involved in each of the comparisons have little in common. It is clear that the linear continuum which is implied in a "more and less" judgment may be conceptual, that it does not necessarily have the physical existence of a yardstick.
>
> And so it is also with attitudes. We do not hesitate to compare them by the "more and less'" type of judgment. We say about a man, for example, that he is more in favor of prohibition than some other, and the judgment conveys its meaning very well with the implication of a linear scale along which people or opinions might be allocated (Thurston & Chave, 1929, p. 10-11).

Figure 2. Learning about attitudes through multiple-choice survey questions involves a double mapping: first, from attitudes to a set, and second, from answer numbers to a number set. Mathematics is used to draw conclusions about the answer numbers. The mathematical conclusions are used to draw empirical conclusions about attitudes. The empirical validity of the mathematical conclusions depends on the validity of both mappings and the soundness of the mathematical logic.



Similarly, they argued for treating attitudes as unidimensional, described an attitude variable as “continuous” and provided an illustrative graph. (Thurston & Chave, 1929). The argument for ordinality is only lightly

evidenced, and no evidence was provided for the other claimed mathematical properties.

Likert, writing in 1932, cited Thurstone and Chave when he assumed that attitudes were on a linear "attitude continuum," which underpinned his explanation of how to build a scale to measure attitudes (Likert, 1932). Likert proposed to measure attitudes by respondent agreement with statements devised by the researcher to mark different points on the attitude continuum. The statements should be arranged in order from one end of the continuum to the other.

Likert then explained that the statements should be mapped to answer numbers, one through five in the case of a five-option question, with the number "one" assigned to one end of the continuum, "three" to undecided and "five" to the other end of the continuum. Likert did not discuss the mathematical properties of these numbers explicitly; however, he recommended calculating the correlation coefficient of the answer numbers of each statement with those of the average score of the battery of statements to ensure that the statement was correctly numbered and offered a table as an example. He treated answer numbers as if they were also real numbers and the attitude continuum as if it were bounded.

Whether the mathematical manipulation of answer numbers leads to valid empirical conclusions depends, among other things, on the validity of the original mathematization of attitudes and the subsequent mathematization of answer numbers. If attitude is ordinal, answer numbers are mathematized as real numbers, and mathematical operations carried out that are only defined for real numbers and not ordinals, then the mathematical operations on the answer numbers have no empirical counterpart and do not provide a foundation for drawing empirical conclusions about attitudes.

If attitudes and perceptions demonstrate the mathematical properties of real numbers, and are bounded, and if the statements offered in a Likert item correctly mark the end points and consistent intervals in an attitude continuum, then there are two possibilities. Answer numbers could be mathematized as ordinals, as the data have the mathematical properties of ordinals, although this results in information loss. Means and standard deviations could not be calculated as they are undefined. Alternately, answer numbers could be real, and the mapping from data to answer numbers is a rescaling. In this case, the answer numbers are analogous to the subject of study, the operations defined, and the mathematical conclusions would lead to valid empirical conclusions.

This is ultimately an empirical question about the nature of attitudes. Revisiting the debate about the mathematical properties of Likert items and scales described earlier, the debate fails to address the mathematical properties of attitudes, on which the proper mathematization of answer numbers depends. In fact, it is not even clear if attitudes are ordered. There is a continuing debate in psychology and economics about whether

evidence shows that preferences demonstrate transitivity (if a > b and b > c, then a > c) (see, e.g., Regenwetter & Dana, 2011; Bleichrodt & Wakker, 2015), a property of both the ordinal and real numbers. Johnson (1936) raised early concerns about whether attitudes are dynamically stable.

Whether various statistical operations on Likert items and scales are defined then depends on how the answer numbers were mathematized. The performance of operations not defined in mathematics is not mathematics and provides no basis for drawing empirical conclusions.

Implications for Quantitative Social Science and Data Science

Stevens' work helped open the door to the use of mathematics to draw empirical conclusions about data that did not have the properties of real numbers. However, it was a product of its time. Stevens built a rickety bridge between the empirical and mathematical worlds and between his data and real numbers, inventing a typology of scales or "levels of measurement." To make his invention work, he offered limitations on the mathematical operations for each scale, but did not explicitly anchor these limitations with reference to mathematics. Moreover, he suggested violating them, which indicated that they were fiat. The flaws in this framework continue to lead to the production of academic work that is not anchored in either the epistemology of science or the logic of mathematics.

Putting the political necessity of establishing that psychology could "measure" aside, Stevens' enterprise is better described as "mathematization" than "measurement." Its purpose was to use mathematics to draw empirical conclusions. It would be more robust if, instead of representing data as "scales" of his own invention, it instead represented them with sets, which are mathematical objects.

If data are mathematized to sets with different or additional mathematical properties, and mathematical operations are performed that depend on those properties, then mathematical conclusions will not yield valid empirical conclusions. However, researchers often begin data analysis by performing mathematical operations defined for the real numbers without investigating the mathematical properties of the subject of study, stating their claims regarding those properties, or providing any evidence of those claims.

The failure to recognize or test such assumptions in psychometrics led Michell (2008) to ask if psychometrics is "pathological science," but the problem is not limited to psychometrics. In political science and economics, many researchers seek to measure social phenomena without examining the mathematical properties of the subject of study. The publication of data sets measuring human rights, corruption, rule of law, democracy, and governance has in turn given rise to large literatures of regression analysis that seek to discover relationships between these and other variables of interest. Such metrics are also used in policymaking. All of these rest on the

unexamined assumption that the underlying phenomena has the necessary properties to be mathematized as real numbers.

The problem is likely to worsen with the advent of data science, as data scientists conduct computational analysis but are often not involved in data collection or decisions about data representation. Not only do they lack access to information about the empirical mathematical properties of the subject of study, the evidence supporting mathematization, and the number set used, but the programming languages they use may or may not permit classification of data by number set or enforce limitations on mathematical operations performed on data based on type. This also encourages the treatment of all numbers as real, reducing the validity of empirical conclusions from mathematical treatment.

Proper mathematization matters for the validity of empirical claims. Accordingly:

1. Researchers should conduct experimental studies to determine if the subject of their study demonstrates mathematical properties such as being well-ordered, transitive, countable, additive, divisible, or continuous; whether it demonstrates closure, commutativity, associativity, or distributivity. The literature on the transitivity of preferences is an example of empirical inquiry regarding one such characteristic. This work is even more strongly needed for variables that are not directly observable, such as attitudes or constructs.
2. Researchers should make their claims about the mathematical properties of the subject of study explicit and support them with empirical evidence.
3. Researchers should make their decisions about mathematization explicit and explain and defend their choices. These are essential premises on which the validity of their results rests.
4. Researchers should mathematize to an analogous number set to ensure the validity of the final empirical conclusion. An analogous number set has only mathematical properties that are demonstrated by the subject of study.
5. Once the data are mathematized to the relevant number set, mathematics determines what operations are defined for that number set. Only defined operations may be conducted, as undefined operations are meaningless in mathematics and do not lead to mathematical conclusions.

One way to see these recommendations is as a prohibition on pseudo-mathematics and a call for more rigor in quantitative social science. But another way is to see them is as a proposal for the opening of programs of empirical study into the mathematical properties of the subjects of research and the exploration of the utility of other number sets. This is fully consistent with Stevens' original vision of using mathematics as a modeling tool.

Author Note: M.A. Thomas, Associate Professor, US Air Force Cyber College.

REFERENCES

Anderson, N. (1961). Scales and statistics: parametric and nonparametric. *Psychological Bulletin, 58*, 305-316.

Bleichrodt, H. & Wakker, P. (2015). Regret theory: A bold alternative to the alternatives. *The Economic Journal, 125*, 493-532.

Boumans, M. (2016). Suppes's outlines of an empirical measurement theory. *Journal of Economic Methodology*, *23*, 305-315.

Burke, C. J. (1953). Additive scales and statistics. *Psychological Review, 60*, 73-75.

Campbell, N.R. (1957). *Foundations of Science: The Philosophy of Theory and Experiment*. Dover.

Carifio, J. & Perla, R. (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*, 106-116.

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

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *American Psychological Society*, *3*, 186-190.

Ferguson, A., Myers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., et al. (1940). The quantitative estimates of sensory events. *The Advancement of Science, 2*, 331-349.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.

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

Johnson, H. M. (1936). Pseudo-mathematics in the mental and social sciences. *American Journal of Psychology 48*, 342-351.

Kero, P. & Lee, D. (2016). Likert is pronounced 'LICK-urt' not 'LIE-kurt' and the data are ordinal not interval. *Journal of Applied Measurement, 17*, 502-509.

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory & Psychology*, 18, 89-109.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 5-55.

Luce, R. & Suppes, P. (2002). Representational Measurement Theory. In H. Pashler & J. Wixted (Eds.), *Stevens' Handbook of Experimental Psychology* (3rd ed., Vol. IV, pp. 1-41). John Wiley & Sons, Inc. https://doi.org/10.1002/0471214426.pas0401.

Luce, R. D. & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27.

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750-751.
Merriam-Webster. Mathematization. Retrieved 12/20/19 from https://www.merriam-webster.com/dictionary/mathematization.
McRae, A. W. (1988). Measurement scales and statistics: What can significance tests tell us about the world? *British Journal of Psychology, 79*, 161-171.
Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 100*, 398-407.
Michell, J. (2005). *Measurement in psychology: A critical history of a methodological concept*. Cambridge: Cambridge University Press.
Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6*, 7-24.
Narens, L. (2002). A meaningful justification for the representational theory of measurement. *Journal of Mathematical Psychology*, *46*, 746–768.
Norman, G. (2010). Likert scales, levels of measurement and the 'laws' of statistics. *Advances in Health Sciences Education, 15*, 625-32.
Petty, N. W. (2013). Oh ordinal data, what do we do with you?" Retrieved from http://learnandteachstatistics.wordpress.com/2013/07/08/ordinal/.
Regenwetter, M. & Dana, J. (2011). Transitivity of preferences. *Psychological Review, 118*, 42-56.
Shen, A. & Vereshchagin, N. K. (2002). *Basic set theory*. Student Mathematical Library, Vol. 17. American Mathematical Society.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.
Thomson, W. (1891). Electrical units of measurement. In Thomson, W. (Ed.), *Popular lectures and addresses. Nature Series. Vol. I. Constitution of matter* (2nd ed., pp. 73-136) London and New York: Macmillan and Co.
Thurstone, L. L. (1929). Theory of attitude measurement. *Psychological Review 36*, 222-241.
Thurstone, L. L. & Chave, E. J. (1930). Theory of Attitude Measurement. In L. L. Thurstone and E. J. Chave (Eds.). (pp. 1–21). *The measurement of attitude*. Chicago: University of Chicago Press.
Trendler, G. (2009). Measurement theory, psychology and the revolution that cannot happen. *Theory & Psychology*, *19*, 579-599.