Measurement Scales and Statistics: A Clash of Paradigms

Joel Michell
University of Sydney
Sydney, New South Wales, Australia

The "permissible statistics" controversy stems from a clash of different theories or paradigms of measurement. Three theories are identified: the representational, the operational, and the classical. In each case the relation between measurement scales and statistical procedures is explored. The representational theory implies a relation between measurement scales and statistics, though not the one mentioned by Stevens or his followers. The operational and classical theories, for different reasons, imply no relation between measurement scales and statistics, contradicting Stevens's prescriptions. A resolution of this issue depends on a critical evaluation of these different theories.

The recent exchange between Gaito (1980) and Townsend and Ashby (1984) shows that the controversy over measurement scales and statistics, begun by Stevens (1946), still persists. The two sides are as unrepentant as ever and show no sign of being able to appreciate the opposing point of view. Such a protracted controversy suggests that the disagreement lies much deeper than the arguments hitherto presented imply. If this is true then merely reciting these well-worn arguments will never resolve the issue. What is needed is a deeper analysis, one that probes to the disagreement's source. This source lies in the different conceptions of measurement. On one side of the debate, those following the Stevens tradition have attempted to make explicit their theory of measurement. This is the *representational* theory. The opponents of this tradition have not been as prepared to lay their cards on the table, but one can discern within psychology at least two other, quite distinct measurement traditions: the *operational* theory and the *classical* theory. The debate on this issue may be advanced by providing a clear statement of each of these theories and their implications regarding the use of statistical (or other numerical) procedures. That is my goal. I will not present arguments for or against these theories at this stage.

Representational Theory and Appropriate Statistics

This theory derives from the writings of Stevens and Suppes (cf. Stevens, 1946, 1951, 1959; Suppes, 1951; Suppes & Zinnes, 1963). They, in turn, were indebted to the earlier representational tradition of Helmholtz (1887), Russell (1903), and Campbell (1920). The core of this theory is that numbers are used in measurement to represent empirical relations between objects. Townsend and Ashby (1984) state clearly this view:

> The fundamental thesis is that measurement is (or should be) a *process of assigning numbers to objects in such a way that interesting qualitative empirical relations among the objects are reflected in the numbers themselves as well as in important properties of the number system.* (p. 394)

It will be helpful to give some simple examples of such "interesting qualitative empirical relations" and the way in which they may be represented numerically. Consider first the representation of an empirical *equivalence* relation. Suppose that people are classified according to hair color. In this case the empirical relation observed is that of Person x's hair being the same color as Person y's. If this relation is transitive, symmetric, and reflexive then it may be represented by the relation of numerical equality (=). That is, numbers may be assigned to the people being classified in such a way that for any pair of people, x and y, x's hair is the same color as y's if and only if $Nx = Ny$ (where Nx is the number assigned to x and Ny is the number assigned to y). For example, blondes may be assigned 1, brunettes 2, redheads 3, and so on. Let us call these assignments "Hair Color Scale X." Scale X is what Stevens would have called a *nominal* scale and, as Suppes and Zinnes (1963) observed, a nominal scale such as X could be replaced by another nominal scale for the same variable (hair color) by any one–one transformation of the numbers assigned. That is, the class of admissible scale transformations for nominal scales is the class of one–one transformations.

Second, consider the representation of an empirical order relation. There are different kinds of order relations, but take for example, a weak order (i.e., a binary relation that is connected and transitive; see e.g., Krantz, Luce, Suppes, & Tversky, 1971). In marking a batch of students' essays the empirical relation observed may be that the quality of Student x's essay is at least as good as the quality of Student y's. If this relation is transitive and connected then it may be represented by the numerical relation of being at least as great as (≥). That is, numbers may be assigned to the essays such that the quality of x's is at least as good as the quality of y's if and only if $Nx \geq Ny$. The result is an *ordinal* scale. For ordinal scales the class of admissible scale

I thank J. P. Sutcliffe, G. W. Oliphant, and two anonymous reviewers for their constructive criticisms of earlier versions of this article.

Correspondence concerning this article should be addressed to Joel Michell, Department of Psychology, The University of Sydney, New South Wales 2006, Sydney, Australia.

Psychological Bulletin, 1986, Vol. 100, No. 3, 398–407

transformations is the class of monotonic increasing transformations.

Third, consider the representation of an empirical order on differences with respect to some attribute. For example, in a psychophysics experiment a subject may be instructed to judge whether or not the difference between a given pair of stimuli, w and x, is at least as great (in some stipulated sense) as the difference between another pair, y and z. If this relation on a set of stimuli satisfies certain testable conditions (viz. those for an infinite difference system, see, e.g., Suppes & Zinnes, 1963), then it may be represented by numerical order on numerical differences. That is, numbers may be assigned to each of the stimuli in the set such that the judged difference between w and x is at least as great as the judged difference between y and z if and only if $N\mathrm{w} - N\mathrm{x} \geq N\mathrm{y} - N\mathrm{z}$ (for all w, x, y, and z in the set). The result is an *interval* scale and the class of admissible scale transformations is the class of positive linear transformations.

Fourth, consider the representation of an order relation on sums (or concatenations) of objects with respect to some attribute. Because psychological examples of such relations are rare, consider the example of length. Let A be a class of rigid, straight rods and for any rods, x and y, in A let $\mathrm{x} \cdot \mathrm{y}$ be the rod obtained by joining x to y, end to end in a straight line. Letting X and Y stand for either single rods in A or concatenations of rods from A, the observed empirical relation is X is at least as long as Y. Providing this relation satisfies the conditions for an extensive structure (cf. Krantz et al., 1971), then it may be represented by numerical order on numbers or sums of numbers. For example, for any w, x, y, and z in A, $\mathrm{w} \cdot \mathrm{x}$ is at least as long as $\mathrm{y} \cdot \mathrm{z}$ if and only if $N\mathrm{w} + N\mathrm{x} \geq N\mathrm{y} + N\mathrm{z}$. The result is a *ratio* scale and the class of admissible scale transformations is the class of positive similarities transformations.

Many of those who support this theory of measurement feel that numbers used to represent one kind of empirical relation (e.g., a mere equivalence relation) cannot always be treated in the same way as numbers used to represent some other kind of empirical relation (e.g., an order on concatenations). Their problem has been to state precisely what this difference in treatment should be.

Stevens thought that relative to each different type of measurement scale (nominal, ordinal, interval, and ratio) certain statistical (or numerical) operations on the numbers assigned were not permissible. For example, for nominal scales the calculation of medians was prohibited and for ordinal scales the calculation of means was prohibited (e.g., Table 6 of Stevens, 1951). His reasoning was that a numerical or statistical operation was not permissible if its result was not invariant under admissible scale transformations. As Adams, Fagot, and Robinson (1965) later showed, Stevens was less than precise about this concept of invariance. That point aside, his emphasis was clearly mistaken. To report the median or mean of a set of measures is to report a fact about them and so it is a bit high-handed to attempt to ban such reports. His opponents could justifiably protest that in science all facts are permissible. Notwithstanding this, however, it is somewhat worrisome when the conclusions derived from measurements depend on quite arbitrary aspects of the chosen measurement scale. So there may have been some point to Stevens's prescriptions. The problem is to express this point in such a way that a researcher's attention to *any* facts concerning his data is not restricted.

Thus, a change of direction came with Suppes (1959) and Suppes and Zinnes (1963). Instead of classifying statistical operations as permissible relative to scale type, it was proposed that measurement statements be classified as *meaningful* or *meaningless*. This approach was developed by Adams et al. (1965) and taken up by Roberts (1979). Townsend and Ashby (1984) treat this approach as if it was the conventional way of handling the problem of the relation between measurement scales and statistics. They do not question its adequacy. However, it has been recognized since its inception that it contains serious difficulties, so it may be worthwhile to mention them again.

There are two versions of the meaningfulness approach and they differ with regard to the kinds of measurement statements they apply to. According to one approach the predicates "meaningful" and "meaningless" are to apply to scale-specific measurement statements, and according to the other they are to apply to scale-free statements. (The terms *scale-specific* and *scale-free* are not actually used in the literature but the intention is plain enough.) By a *scale-specific* measurement statement I mean one containing metrical predicates that include reference to a particular scale of measurement. Some examples follow.

1. The sum of the Scale X hair color measures for Sample M is 10.
2. The mean Mohs' scale of hardness measure for minerals in Sample T is greater than the mean Mohs' scale of hardness measure for minerals in Sample R.
3. Today's temperature in degrees centigrade is twice yesterday's temperature in degrees centigrade.
4. The average height in feet of red kangaroos is 5.3.

Each of these statements is scale specific because within each the variable referred to (hair color, hardness, temperature, and height) is described relative to a particular measurement scale (Scale X, Mohs' scale, degrees centigrade, and feet).

Suppes and Zinnes (1963) and Roberts (1979) each present a criterion of meaningfulness that is intended to apply to what I call scale-specific statements. Suppes and Zinnes state, "A numerical statement is *meaningful* if and only if its truth (or falsity) is constant under admissible scale transformations of any of its numerical assignments, that is any of its numerical functions expressing the results of measurement" (p. 66). Roberts's criterion is similar: "A statement involving numerical scales is meaningful if and only if its truth (or falsity) remains unchanged under all admissible transformations of all the scales involved" (p. 71). Fortunately, the statement of both criteria is supplemented by examples showing explicitly what is meant. Consider an example given by Roberts (1979).

> Let us first consider the statement
>
> $$f(a) = 2f(b),$$
>
> where $f(a)$ is some quantity assigned to a, for example, its mass or its temperature. We ask under what circumstances this statement is meaningful. According to the definition, it is meaningful if and only if its truth value is preserved under all admissible transformations ϕ, that is, if and only if, under all such ϕ,
>
> $$f(a) = 2f(b) \leftrightarrows (\phi \cdot f)(a) = 2[(\phi \cdot f)(b)]. \text{ (p. 71).}$$

(Here, of course, "$\phi \cdot f$" is understood as the composition of the two functions ϕ and f.) This example shows that the meaningfulness criterion is intended for application to scale-specific statements, for the statement $f(a) = 2f(b)$ is of a kind with Statement 3. That is, the terms $f(a)$ and $f(b)$ denote actual measurements. Now this observation reveals a certain incoherence in the way the criterion is expressed. What is being considered is not the truth value of a single scale-specific statement under admissible transformations of the scale values involved but, rather, the truth values of an infinite class of scale-specific statements, one for each different admissible transformation. That is, to use Roberts's terminology, for each distinct admissible transformation, ϕ, $(\phi \cdot f)(a) = 2[(\phi \cdot f)(b)]$ will be a different statement. For example, if $f(a)$ is the temperature of a in degrees centigrade, then if ϕ is the function °F = 1.8 °C + 32, the statement $(\phi \cdot f)(a) = 2[(\phi \cdot f)(b)]$ becomes

3a. Today's temperature in degrees Fahrenheit is twice yesterday's temperature in degrees Fahrenheit.

Yet if ϕ is some other admissible scale transformation (in this case any other positive linear transformation) then a quite different statement results. This statement will be analogous to (3) except that it will be made relative to some possible scale for the measurement of temperature other than degrees centigrade (and, of course, degrees Fahrenheit). Because there are an infinite number of such transformations there will be an infinite number of such statements analogous to (3). So what is really being claimed by these authors is that a statement like (3) is meaningless, not because its truth value is not invariant under admissible scale transformations, but because its truth value is not the same as that of each of these statements analogous to (3). If this point is sharpened then the intention of the proposed meaningfulness criterion can be stated more precisely.

Let s be any scale-specific statement made relative to Scale f. Corresponding to s is a family of analogous scale-specific statements, S, such that any statement, s′, belongs to S if and only if s′ is identical to s except that where s makes mention of Scale f, s′ makes mention of Scale g (where g and f are scales for the measurement of the same variable and g is related to f by some admissible scale transformation). Let us call S the family of admissible transformations of s. Now, what Suppes and Zinnes (1963) and Roberts (1979) really mean is that a scale-specific statement s is meaningful if and only if each member of S (the family of admissible transformations of s) has the same truth value as s. Thus (3) is meaningless because (3a) belongs to its family of admissible transformations and (3a) is false whenever (3) is true.

Having stated precisely what is meant, the criterion can now be evaluated. It would be fatuous to object that the criterion fails because a statement must be meaningful in order to have a truth value in the first place. Obviously, these authors do not mean by *meaningful* "possessing meaning," and by *meaningless*, "possessing no meaning." Perhaps they should have chosen a different term here, as did Adams et al. (1965) who used "scientific significance" instead. Whatever terms are used, however, the intention is to classify scale-specific statements into two categories, one consisting of statements whose truth (or falsity) is an artifact of the particular measurement scale chosen (the "meaningless" statements) and the other consisting of statements whose truth (or falsity) is not an artifact of the measurement scale chosen (the "meaningful" statements). On top of that purely descriptive function a prescriptive intention is conveyed by the use of the commendatory term "meaningful" and the pejorative term "meaningless." This is the connotation that one should confine attention to the meaningful and abstain from using the meaningless. The difficulty for this approach is that "meaningless" statements are scientifically useful and, hence, not necessarily to be abstained from.

This point was noticed by Adams et al. (1965) and emphasized by Adams (1966). They pointed out that all statements reporting the results of individual measurements (e.g., the weight in pounds of Object X is 10) fail to pass the criterion and, hence, are meaningless. As Adams (1966) noted, "If such statements were excluded on the grounds that they are not 'meaningful,' all *data* would be banished from science" (p. 132).

It is tempting to overlook this problem because the meaningfulness criterion was never intended to evaluate such basic measurement statements. Yet this example merely signals a difficulty that is more deep-rooted. Suppose that Hair Color Scale X is used to measure the hair colors of Sample M of subjects and it is found that Statement 1 is true. Of course Statement 1 is "meaningless" because there will be some one-to-one transformation of Scale X values for which the sum of the scale values for Sample M is not 10 (e.g., if the transformation involves making each scale value greater than 10). Yet despite being "meaningless," Statement 1 is scientifically useful in the sense that it implies true consequences about the nature of Sample M. For example, from (1) it follows that

5. Not all of the members of sample M are redheads.

This is not an isolated example. There are many others. Another relates to Statement 2. Statement 2 implies that at least one of the minerals in Sample T is harder than some mineral in Sample R and this implication holds even if (2) is meaningless relative to this criterion (which depends on how the hardnesses of the minerals are distributed over the two samples). Or, to take another example, Statement 3, which has already been shown to fail this criterion, implies that today's temperature is greater than yesterday's. The point is that "meaningless" scale-specific statements sometimes entail true empirical consequences and consequently are of scientific use or significance. This observation should cause us to question the prescriptive force of the distinction between "meaningful" and "meaningless" scale-specific measurement statements.

For reasons like this Adams et al. (1965) shifted the focus of the meaningfulness approach from scale-specific to scale-free measurement statements. Each scale-specific statement possesses a scale-free version. The scale-free version of a scale-specific statement, s, is a statement, r, that is identical to s in all respects except that terms in s refering to particular measurement scales are dropped. Thus, the scale-free versions of Statements 1–4 are Statements 1′–4′.

1′. The sum of the hair colors for Sample M is 10.

2′. The mean hardness of minerals in Sample T is greater than the mean hardness of minerals in Sample R.

3′. Today's temperature is twice yesterday's.

4′. The average height of red kangaroos is 5.3.

It can be argued that one's aim in making measurements is not to derive scale-specific results but scale-free results. For example, in comparing the heights of men and women one does not want to know that

6. The average height of men in inches is greater than the average height of women in inches (a scale-specific statement),

except insofar as it enables one to know that

6′. The average height of men is greater than the average height of women (a scale-free statement).

So there is some force to the suggestion that the "meaningfulness" of scale-free statements is of primary interest.

The criterion that Adams et al. (1965) suggested amounts to saying that a scale-free measurement statement is meaningful if and only if the truth values of all of its scale-specific versions are the same. (A scale-specific version of a scale-free measurement statement, r, is any statement, s, identical to r except that each [measurable] variable named in r is described in s by reference to a particular scale of measurement, the same scale being mentioned for all references to the same variable. Of course, each member of the admissible transformations of s [i.e., each member of S] will be a scale-specific version of r and, hence, there must be an infinite number of them). So that, for example, Statement 4′ is "meaningless" because given the truth of (4), it follows that another scale-specific version, (4a), must be false and vice versa, where (4a) is

4a. The average height in meters of red kangaroos is 5.3.

Adams et al. (1965) insist that by *meaningful* they mean semantically meaningful (or having a truth value) and that by *meaningless* they mean semantically meaningless (or not having a truth value). This seems to be an overstatement. For example, Statement 4′ does have a truth value: it is simply false. The average height of red kangaroos must be some particular height and a height is a (possible) property of some object and not a number. Statement 4′, however, asserts it to be a rational number rather than a particular height. Hence, what Statement 4′ asserts is false. Yet this does not matter, for this criterion, like the last, is intended to discourage us from using so-called "meaningless" statements. These statements are really those in which the consequences of arbitrary scale features are taken to be true of the underlying variable itself. Statement 4′, for example, treats an average of height measurements as if it was a value of the height variable and not simply a result contingent on using the foot scale. Each of Statements 1′–3′ commits a similar error. Statement 6′, on the other hand, does not. It generalizes from height measurements (Statement 6) to heights themselves, a result that is not a mere artifact of using the inch scale. Clearly, the distinction between "meaningless" and "meaningful" scale-free statements is important. However, it has not been adequately captured, for as it stands it encounters difficulties similar to those of the previous criterion.

When Suppes (1959) first sought a solution along these lines he noted a significant problem. Statements such as

7. Smith's height is 6.4

and

8. Jones's height is 3.2

fail this criterion (i.e., they are "meaningless" scale-free statements). However, they jointly entail that

9. Smith's height is twice Jones's

and (9), of course, is "meaningful" and may indeed be true. So once again we are confronted by the problem that statements judged "meaningless" entail empirical statements that may be true and of some scientific significance.

The fact that Stevens with his prescriptions about permissible statistics and later Suppes, Zinnes, Adams, Fagot, Robinson, and Roberts with their criteria of "meaningfulness," gained a considerable following among supporters of the representational theory suggests that they were close to proposing a satisfactory statement of the relation between measurement scales and statistics. Unfortunately, they have not quite succeeded, and in order to see where they have gone astray it is helpful to look again at the underlying logic of the representational theory of measurement.

The central principle of this theory is that measurement is the numerical representation of empirical facts. Within this theory there have been internal debates about just what kinds of empirical facts measurements represent. For example, Campbell (1920) insisted that measurement should be the numerical representation of facts about concatenations, or at least based in some way on such facts. In Stevens's terminology this would limit measurement to ratio scaling. Yet as Stevens (1951) pointed out, in so restricting the concept, Campbell was not adhering to the central principle of representationalism. Russell (1903) had earlier included the numerical representation of ordinal structures within the concept of measurement and Stevens was even more liberal in allowing measurement to include the numerical representation of merely classificatory structures. These controversies were simply the growing pains encountered as the representational theory freed itself from the classical theory of measurement and followed the internal logic of its central principle as applied to the subject matter of psychology. According to this theory, the numerical representation of any empirical structure is measurement.

Given this sketch of representationalism the obvious question is "Why assign numbers to represent empirical structures?" Campbell and Russell had no doubts about the answer to this question. It was, said Campbell (1920), so that "the powerful weapon of mathematical analysis" could "be applied to the subject matter of science" (pp. 267–268). Mathematical analysis is powerful because it contains a storehouse of valid argument forms or theorems that may be applied to empirical propositions once numerical assignments are made. This enables us to derive empirical conclusions from data via mathematical arguments. In this way such conclusions could be drawn more conveniently. As Russell (1983) noted in one of his earliest papers: "Number is of all conceptions, the easiest to operate with, and science seeks everywhere for an opportunity to apply it" (p. 301). What needs to be stressed, however, is that the conclusions reached via numerical argument must be conclusions that are wholly implied by the empirical data itself and not conclusions whose content depends on the numbers assigned. Otherwise measurement would be more than mere numerical representation and the function of numbers would be more than the mere lubrication of the deductive process. Field (1980) has aptly put the matter thus: "The conclusions we arrive at by these means are not new, they are already derivable in a more long-winded fashion from the premises without resource to mathematical entities" (pp. 10–11).

Because the numbers used in measurement are a mere convenience and cannot contribute content to the conclusions derived and because these conclusions must be already entailed by nonnumerical, empirical data (albeit, long-windedly), it follows that they cannot be scale-specific statements and so must be scale-free statements. Putting the matter this way it becomes obvious that what is of interest about any scale-free statement made as a result of measurement is simply whether or not it really does follow validly from the empirical observations underlying the measurements. These observations will be statements about which empirical objects stand in which empirical relations to which other empirical objects. I call these observations the *scale-free premises.* The scale-free premises are represented numerically through measurement. The measurements lead first to scale-specific conclusions and then from these, scale-free conclusions are inferred. The relevant issue here is whether or not these scale-free conclusions really do follow from the scale-free premises. The question never was one of permissible statistics or of meaningfulness. It was always only one of legitimate inference.

This resolution of the problem may be given a more exact formulation. According to the representational theory, measurement begins with the identification of a set of empirical objects and empirical relations between these objects. Let the set of objects be A and the relations between them be $R_1, R_2, \ldots, R_n$ (where n is an integer ≥ 1). (In each of the examples mentioned earlier in this section only a single empirical relation was mentioned [i.e., equivalence, order, order on differences, and order on concatenations] relative to each measurement scale but n may exceed 1.) The empirical structure to be represented numerically is then A together with $R_1, \ldots, R_n$. Following Suppes and Zinnes (1963), we may think of this empirical relational system as an ordered set, $\langle A, R_1, \ldots, R_n \rangle$. Any statement describing some fact about the structure of an empirical relational system is a scale-free premise relative to that system.

For example, an empirical relational system for the measurement of weight might consist of a set, A, of marbles of various weights and two relations determined by operations on a beam balance. One of these, R_1, may be a binary weak order, such that any element of A, a, stands in R_1 to any other element, b, if and only if a is at least as heavy as b (which may be determined by placing a alone in one balance pan, b alone in the other, and noting the outcome). The other relation, R_2, may be a ternary relation such that any pair of elements of A, a and b, stand in R_2 to c if and only if a and b together are at least as heavy as c (which may be determined by placing a and b in one balance pan, just c in the other, and noting the outcome). Not only may the number of relations in an empirical relational system exceed one but, as shown in this example, they need not necessarily all be binary relations. They may be ternary, quarternary, or of any finite order.

An empirical relational system, $\langle A, R_1, \ldots, R_n \rangle$ is represented numerically by finding a set of numbers, N, and a set of n numerical relations, $S_1, \ldots, S_n$, such that $\langle a_1, \ldots, a_m \rangle$ is any m termed ordered sequence of elements of A standing in relation R_i (for any $i = 1, \ldots, n$) if and only if $\langle n_{a_1}, \ldots, n_{a_m} \rangle$ is an m termed sequence of elements of N standing in the relation S_i (where $n_{a_i}, \ldots, n_{a_m}$ are the numbers assigned to $a_1, \ldots, a_m$, respectively). That is, each empirical object is assigned a number so that each empirical relation is represented by a numerical relation. Any statement describing the numbers assigned to objects is scale-specific and from such scale-specific statements others may be deduced by numerical argument forms (or calculations). At some point in this process one may arrive at a scale-specific statement (e.g., Statements 1–4 and 6) from which one wants to infer the scale-free version (i.e., Statements 1′–4′ and 6′). Under what conditions is such an inference valid?

The most general answer to this question is that the scale-free version of a scale-specific statement follows validly from it if and only if that scale-free statement follows validly from the scale-free premises by some chain of nonnumerical argument. This general answer, however, is of no help in particular cases, as its application requires carrying out the long-winded, nonnumerical chain of argument that measurement is intended to circumvent. A criterion is needed that can be applied directly to either scale-specific or scale-free conclusions.

By this stage it must be obvious that what has hitherto been called the "meaningfulness criterion" is, in fact, a necessary condition for valid argument from scale-specific statements to their scale-free versions. Let s be any true scale-specific statement and let r be its scale-free version. Then r follows validly from s only if each member of the family of admissible transformations of s (i.e., each s′ in S) is also true. This criterion ensures that the fact that r follows from s does not in any way logically depend on the use of any particular scale of measurement. Had another scale (related to the chosen one by some admissible transformation) been used, the same scale-free conclusion would have resulted. Anything less than this condition could allow the inference of contradictory scale-free conclusions from the same set of data.

For example, suppose that

10. The height in meters of Smith times the height in meters of Jones equals the height in meters of Brown.

Yet we cannot validly infer from (10) that

11. The height of Smith times the height of Jones equals the height of Brown.

The reason is that if (10) is true then (12) is false.

12. The height in feet of Smith times the height in feet of Jones equals the height in feet of Brown.

If (11) follows from (10) then not (11) follows from not (12). Therefore, (11) cannot follow from (10) without contradiction. This necessary condition for valid inference rules out the possibility of such contradictory inferences from the same body of data.

Although this condition is necessary for the valid inference of scale-free conclusions from their scale-specific versions, it is not sufficient. For certain sets of data a conclusion may satisfy this condition but not follow from the scale-free premises. For example, let T and R be two sets of minerals such that each mineral in T is harder than any mineral in R. Then, if these minerals are assigned ordinal scale values to represent their hardnesses on, for example, Mohs' scale of hardness, Statement 2 will be true. What is more, the family of admissible transformations of (2) will also be true. Yet Statement 2′ cannot validly follow from (2) because (2′) does not follow from the set of scale-free premises in this case. In the case of an ordinal scale the scale-free premises will simply be statements describing order relations between objects, like (13).

13. Mineral x is at least as hard as Mineral y.

Statement 2′ concerns mean hardnesses. The concept of a mean is not definable in terms of order relations alone. Hence, from a set of statements all of the form of Statement 13 (reporting hardness order relations between objects in Sets R and T) the scale-free conclusion (2′) cannot follow. Of course, if it was known that hardness possessed a structure that was richer than mere order, then (2′) might follow, but it does not follow validly from the ordinal information alone.

Thus, a second necessary condition for valid argument from a scale-specific statement, s, to its scale-free version, r, is that all concepts within r be definable in terms of $R_1, \ldots, R_n$, the relations involved within the relevant empirical relational system. Of course, this condition needs to be worked out in detail for particular kinds of concepts relative to particular kinds of empirical relational systems (e.g., product–moment correlation coefficients relative to purely ordinal systems, etc.). Yet that is a difficult task and is best left to those who accept this theory of measurement.

These two necessary conditions for legitimate inference capture the underlying motivation for Stevens's prescriptions about permissible statistics and the attempts by Suppes and others to formulate a criterion of "meaningfulness." The rules and criteria they laid down make some sense in the light of the abovementioned considerations.

However, that is not the full story of the relation between measurement scales and statistics within the representational theory. There appear to be cases where a scale-free statement, r, validly follows from a scale-specific statement, s, when r is not the scale-free version of s. The deduction of (5) from (1) is a case in point. As already noted, not all permissible transformations of (1) will be true when (1) is true, so there is no question of validly inferring (1′) from it. Be that as it may, (5) follows from it and the inference is obviously valid. (Actually [5] follows not from [1] alone but from [1] together with a description of Scale X.) Thus what we have are necessary conditions for the valid inference of scale-free conclusions from their scale-specific versions and not a general criterion for the valid inference of scale-free statements from scale-specific statements. Clearly, the issue is more complex than Townsend and Ashby (1984) would have us believe.

Operational Theory and Appropriate Statistics

Attractive as the representational theory is, it has failed to gain a universal following within psychology because a number of cherished psychometric methods apparently do not conform to it. In particular, it is not clear in the case of mental tests and summated rating scales (which account for a larger proportion of quantitative activity within psychology) exactly what empirical relations are being represented. For example, consider mental tests in more detail. In the most common case of tests composed entirely of dichotomous items, for each person doing the test the data consists of an ordered sequence of responses, each response being classified as "correct" or "incorrect." One empirical relation of interest within this data is that of one person, i, getting correct at least all those test items that another person, j, gets correct. In this case, i's performance on the test is at least as good as j's. If this relation is transitive and connected over the members of some population, then it constitutes a weak ordering of them (commonly referred to as a Guttman scale, cf. Guttman, 1944). Such a weak ordering may be represented numerically, thus producing an ordinal scale of measurement according to the representational theory. However, applying mental tests to random samples from the populations for which they are intended rarely, if ever, produces Guttman scales, because the abovementioned relation fails to be connected. Hence, according to the representational theory, mental tests provide something less than an ordinal scale. Most psychologists are loathe to accept this conclusion. The standard psychometric practice is to treat the person's number of correct responses to the test items as a measurement of some kind. This practice does not conform to the requirements of the representational theory because it is not obvious just what "interesting qualitative empirical relations" are represented by such total test scores. Of course, these procedures may not be measurement procedures at all, but this is clearly not the opinion of the substantial number of psychologists who use them.

The interesting thing is that although the representational theory itself is not endorsed by all psychologists, Stevens's well-known definition of measurement as "the assignment of numerals to objects or events according to rules" (1946, p. 667, 1951, p. 1, and 1959, p. 19) is almost universally accepted. There is no contradiction here, for Stevens's definition is wider than the representational theory alone requires. This theory requires that the numerical assignment rules be limited to those in which the numbers assigned represent empirical information, but Stevens's definition does not contain that limitation. Indeed, Stevens actively encouraged a wider interpretation. For example, he once added that "provided a consistent rule is followed, some form of measurement is achieved" (1959, p. 19). Although this sits uneasily with his representationalism, it must be remembered that in his theory he attempted to weld together two measurement traditions: representationalism and operationalism.

The second of these traditions had already received warm acceptance within psychology when Stevens first expounded his theory and it was this strain in his theory that was most acceptable to the majority of psychologists. His representationalism, with its alleged implications about permissible statistics, was strongly resisted by many, but his operationism, as reflected in his definition of measurement, was liberally interpreted within psychology. It is still so. Consider Cliff's (1982) defense of the rating scale method: "Crude as they are, rating scales constitute a workable measurement technology because there has been repeated observation that numbers assigned in this way display the appropriate kinds of consistency" (p. 31). This defense is operationist in spirit and a large number of psychologists would be sympathetic toward it.

As is well known, operationism derives from the methodological writings of Bridgman (1927) and it is aptly summed up in one of his slogans: "In general we mean by any concept nothing more than a set of operations; *the concept is synonymous with the corresponding set of operations*" (p. 5). Because the process of measurement is always an operation of some kind on the object to be measured and because it gives rise to numerical results, it would seem that from the operational point of view, measurement is simply an operation that produces numbers or numerals. This understanding is very close to Stevens's definition, for assigning a numeral according to a rule is simply a kind of operation and the numeral assigned is the outcome of the operation. A similar interpretation of Bridgman's doctrine was made by Dingle (1950) in his exposition of an operational theory of measurement. He defined measurement as "any precisely specified operation that yields a number" (1950, p. 11).

The fundamental difference between the operational and representational theories turns on this point of how numbers get into measurement. They both agree that measurement involves making numerical assignments to things. However, according to the representational theory, the numbers represent an empirical relational system, which is thought of as an objective structure existing quite independent of our operations. Numbers are used as a convenience and are, in principle, dispensible. This is not so, according to operationism. According to it numbers do not point beyond themselves to a scale-free realm. Rather the data on which measurement is based are inherently numerical. They are numerical because the operations involved produce numbers. For the strict operationist, science is simply the study of our operations and not the study of a reality that is thought to lie beyond them.

So, for example, the operationist considers test scores as measurements simply because they are reasonably consistent numerical assignments that result from a precisely specified operation. This information is not sufficient for the representationalist, however. If test scores are to be counted as measurements in their view then the numerical relations between them (e.g., one score being greater than another, etc.) must represent qualitative, empirical relations between test performances and research must be devoted to identifying such relations and describing their properties. Thus, in the attempt to produce measurement, operationists and representationalists will have quite different research interests. The operationist will be interested in devising operations that produce reasonably consistent numerical assignments. The representationalist will be interested in finding empirical relations that display properties similar to those of relations between numbers (e.g., orders, orders on differences, etc.).

According to the operational view the aims of quantitative science are quite simple. Given operations for making consistent numerical assignments, the aim is to discover quantitative relations between them. It is in this context that the concept of scale type may find application. For example, a psychologist who takes scores on the Wechsler Adult Intelligence Scale (WAIS) as definitive of adult intelligence may regard scores on another test as an ordinal scale of intelligence if they are nonlinearly but monotonically related to WAIS scores (within the limits of error). On the other hand, if they are linearly related to WAIS scores then they may be regarded as an interval scale of intelligence. This example shows that in the operational theory, scale type is not relative to the kind of empirical relation represented (as in the representational theory) but, instead, to the kind of use one wants to make of the numerical information in the measurements. If one is doing no more than classifying based on the numbers assigned then one has a nominal scale; if one uses no more than ordinal information then one has an ordinal scale; if one is using information about differences then one has an interval scale; and if one is using information about ratios then one has a ratio scale. In this conception of scale type, numbers assigned to objects "do not know where they came from" (Lord, 1953, p. 751) and so we are free to use them as we please.

That is, in the operational view of measurement, there can be no restrictions on statistical or other numerical procedures relative to scale type. For example, an ordinal scale of intelligence is not an ordinal scale in any absolute sense and a researcher is not restricted in how much scale values may be treated numerically. Numerical (or statistical) results based on measurements are seen as an end in themselves and not as a stage along the way to scale-free conclusions. To a supporter of this view, Stevens's prohibitions and the later meaningfulness criteria would be anathema. This is because they stand in the way of the discovery of quantitative relations between the outcomes of different operations and, hence, would be seen as an obstacle to scientific progress. Little wonder that many psychologists strongly resisted Stevens's prescriptions.

Some psychologists, although sympathetic to the empiricist spirit behind operationalism, do not accept this definition of measurement. They prefer a narrower concept. They argue that test scores, rating scales, and other outcomes of standard psychometric procedures reflect the structure of underlying theoretical variables, variables that are not themselves directly measurable. For example, it is often claimed that scores on cognitive tests reflect levels of latent abilities. Such an approach to psychological measurement owes more to the classical theory of measurement than to either the operational or the representational theories.

Classical Theory and Appropriate Statistics

It would be a mistake to think that the representational and operational theories of measurement are the only ones adhered to by psychologists. Yet this presumption is made by many modern expositors of the theory of measurement, for they write as if inquiry in this area began with Helmholtz (1887) and Russell (1903). Some (e.g., Ellis, 1966) refer to an earlier view, but always only in order to dismiss it cheaply; never in order to understand it. Traces of this earlier view may be found in the works of Aristotle (see Apostle, 1952) and Euclid (see Heath, 1908) and no doubt exist even prior to that. This view of measurement was further developed during the Middle Ages and the Scientific Revolution and sustained the practice of measurement until at least the beginning of this century. It was in the context of this theory that modern psychology was conceived of as being a quantitative science. We find it presumed in Fechner's (1860/1966) claim that "generally, the measurement of a quantity consists of ascertaining how often a unit quantity of the same kind is contained in it" (p. 38), and Titchener's (1905) remark that "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" (p. xix). Though now largely supplanted in the minds of psychologists by the representational and operational theories, vestiges of the classical theory still persist in psychology (see, e.g., Jones, 1971; Rozeboom, 1966).

According to this theory, measurement is "the assessment of quantity" (Rozeboom, 1966, p. 224), that is, the assessment of "how much." In measurement one is concerned with how much of a given attribute some object possesses (e.g., how much mass, intelligence, etc.). The assessment of quantity only applies to quantitative variables or, as Jones (1971) puts it, "To be measurable an attribute must fit the specifications of a quantitative variable" (p. 336). Attributes like mass can be measured because they are quantitative, but attributes like nationality cannot because they are not quantitative. The difference between quantitative and nonquantitative attributes resides in the structure of the attribute itself. Different masses may be ordered and have additive relations to one another; different nationalities do not.

The precise specification of the structure of a quantitative attribute is a matter of some complexity and would be out of place here. Suffice it to say that the values of such an attribute stand in ordinal and additive relations to one another so that they form a structure similar to that described by Krantz et al. (1971) as an *extensive structure.* There is, however, a fundamental difference between their concept of an extensive structure and that contained within the classical theory. They regard the elements that possess an extensive structure as *objects* (e.g., straight, rigid rods), whereas according to the classical theory they are *attributes of objects* (i.e., the lengths of such rods). Stated this way the difference appears to be minor but, in fact, it is of the first importance. The attribute itself is taken to be extensive (or quantitative) and not necessarily the objects possessing it. Thus, the fact that an attribute is quantitative in no way depends on the existence of a set of objects possessing a relation of physical addition (or concatenation) between them.

Hence, it is no problem for the classical theory that an attribute like temperature is, within modern physics, thought to be quantitative, even though objects possessing temperature do not form an extensive structure. This is an important difference between the classical and the representational theories. This latter theory requires that the relations between the relevant objects constitute an appropriate kind of empirical relational system. Only then is measurement possible. The classical theory requires no such thing. It merely requires evidence supporting the hypothesis that the attribute to be measured is quantitative, but no limitation is put on the form this evidence must take. It may, as with length, amount to the discovery that certain objects possessing length constitute an empirical relational system of an extensive kind, or it may, as with temperature, be much less direct. In either case, the important thing is that the hypothesis that a given attribute is quantitative be supported in some way by observational evidence and that there be no falsifying evidence. Such a hypothesis is, in this respect, no different from any other hypothesis in science and one cannot lay down in advance what will or will not count as evidence for or against it. That is a matter that must be left to the ingenuity of the researchers.

Given that an attribute is quantitative, it follows that it is measurable. That is, taking some value of the attribute as the unit of measurement, the numerical relation between the unit and the value to be assessed may be determined or approximated. This is also an important point, for according to the classical theory numbers are not "assigned" in measurement, rather, numerical relations are discovered. Any two values of the same quantitative attribute will stand in a relation of relative magnitude to one another and this relation is numerical. Take, for example, the length of my arm and the length of my big toe. It is simply a fact that

14. The length of my arm is 18.15 times the length of my big toe.

More generally, any length is twice, thrice, or *n* times (where *n* is some positive real number) any other length and so can be identified as *n* times the unit. Thus, according to the classical theory such numerical relations as one length being *n* times another are empirical relations between lengths, and measurement is the discovery of such relations. This aspect of the theory has been placed on a firm footing by Bostock (1979), who has shown how ratios of values of continuous quantitative attributes constitute a complete ordered field and consequently can be taken to be the real numbers (for that which has the structure of the real numbers is the real numbers).

The classical theory differs from both the representational and the operational theories in holding that in measurement numbers (or numerals) are not assigned to objects, rather numerical relations between values of a quantitative attribute are discovered. It differs from the representational theory as well in not specifying the kinds of observations necessary for measurement. Yet it differs from the operational theory in making the possibility of measurement a matter of evidence, rather than simply a matter of constructing number generating operations.

The classical theory sustained the development of most quantitative theories in psychology prior to the 1950s. In the factor analytic theories of ability proposed by Spearman (1927) and Thurstone (1938), in the Thurstonian theories for psychophysical and attitude measurement (Thurstone, 1927a, 1927b), and in Hull's (1943) theory of learning (to mention just a few), quantitative attributes of various psychological kinds were hypothesized. In none of these theories could the psychological variables be measured directly. The development of suitable measurement procedures followed neither the representational nor the operational pattern. That is, no attempt was made to specify empirical relational systems and measurement was not seen to be simply a matter of devising number generating operations. Instead, the theory surrounding the hypothesized psychological attributes was elaborated in such a way as to relate them to observable quantities of one kind or another (e.g., test scores, relative frequencies or response speeds, etc.). In this manner the development of quantitative psychology conformed to the classical theory and followed the path of quantitative physics. The legitimacy of the measurement procedures so constructed was taken to be contingent on the truth of the underlying psychological theories, for the hypothesized quantitative attributes were part and parcel of such theories. It was only with the increasing acceptance of operational views that these and other measurement procedures were seen as establishing psychological measurement independent of the underlying theories. Thus, for example, many accepted that mental tests measured intelligence quite independent of the truth of theories like Spearman's or Thurstone's. This change of attitude marked the switch from the classical to the operational theory.

Because, according to the classical theory, measurements are always real numbers, it follows that any valid numerical argument form may be applied to them. That is, for this theory all measurements are of the same scale type and no restrictions of the kind proposed by Stevens on statistical operations apply.

Obviously, the measurements that Stevens and Suppes call ratio scale measurements fit this theory. So does their category of interval scale measurements. This is because differences within an interval scale constitute a ratio scale. What is being measured then (in the classical sense) with an interval scale are really differences. For example, the Celsius scale of temperature measures—not temperature (as such) but temperature differences—the unit of measure (1 °C) is itself a difference (i.e., one-hundredth part of the difference between the boiling and freezing points of water). Thus scales like this within the representational theory fit the classical pattern. Yet so do many that do not conform to the representational theory; for example, factor scores in ability measurement and Thurstone scale values in attitude measurement.

The status of Stevens's ordinal scale type is not so clear. Yet not everything thought of as an ordinal scale necessarily is one. For example, Townsend and Ashby (1984) take it for granted that rating scales are merely ordinal and, perhaps, one would be hard-pressed to demonstrate that they carry more than ordinal information. Nevertheless, a researcher using such a scale may take it that the attribute being rated (say, degree of agreement with some attitude statement) is a continuous quantity and that subjects can judge numerical relations on this attribute to a certain (perhaps, rough) degree of precision. Thus, such a researcher will take these ratings to be quantitative information in the full-bodied, classical sense (though they may also be seen as containing a high degree of error). The objection that such a researcher's assumptions may be false is true of course, but irrelevant in this context. The point is that in using statistical procedures that leave Stevens's followers outraged, a researcher may be interpreting measurements within the classical theory, and so may be acting quite properly according to this theory. The issue of whether what are taken to be measurements (in the classical sense) in a given field really are measurements is an issue that must be faced. Yet science often makes progress by only later finding ingenious and novel ways to justify earlier assumptions (as Krantz, 1972, has done for magnitude estimation). The important thing is not to discourage speculation, but rather to recognize when one is skating on thin ice.

As for Stevens's nominal and ordinal scales in the strict sense, they are much more accurately described as numerical coding than as measurement. In these cases numerals are used to label nonnumerical properties or relations. It is a harmless device as long as it is clearly distinguished from measurement. The fundamental difference is that in measurement, numerals are used to refer to numerical relations; in numerical coding they are not. Numerical coding finds a parallel in the use of color terms by physicists to label properties of quarks. One, of course, would no more think of applying the truths of arithmetic to numerical labels so used than one would think of applying the laws of color mixture to those properties of quarks labeled with color words.

So according to this theory, some of the variables studied in science are quantitative and measurement is the discovery of certain kinds of numerical relations (viz. ratios) among the values of such variables. The numbers (numerical relations) so discovered are real numbers and are as empirical as any relations found in scientific research. The full range of valid numerical argument forms may be legitimately applied to them. What is more, the numerical conclusions so obtained are directly interpretable as empirical claims, be they scale specific or scale free. In this respect the classical theory differs sharply from the representational theory.

Summary

Three theories of measurement exist side-by-side within psychology. Each theory has different implications for the relation between measurement scales and statistics. Failure to recognize this has hindered progress over the past three decades on what has become known as the "permissible statistics" controversy. The supporters of the representational theory presumed that they knew what measurement really was and they have taken it on themselves to prescribe some aspects of statistical practice within psychology. This presumption has not gone unchallenged. The representational theory is not the only theory of measurement. There are at least two others, the operational and the classical.

When looked at closely the representational theory entails no prescriptions about the use of statistical procedures (as Stevens thought), nor does it imply anything about the meaningfulness of measurement statements (as current belief has it). It does, however, imply something about the conditions under which scale-free conclusions follow from their scale-specific versions.

The operational theory implies that there is no relation between measurement scales and appropriate statistical procedures. The classical theory rejects Stevens's scale-type distinctions and does not prohibit the use of any statistical procedures with measurements.

Thus, it can be seen that the issue is not as simple as the representationalists have argued. Only after these three theories have been critically examined and one is shown to be preferable to the others will the issue be resolved.

References

Adams, E. W. (1966). On the nature and purpose of measurement. *Synthese, 16,* 125–169.

Adams, E. W., Fagot, R. F., & Robinson, R. E. (1965). A theory of appropriate statistics. *Psychometrika, 30,* 99–127.

Apostle, H. G. (1952). *Aristotle's philosophy of mathematics.* Chicago: University of Chicago Press.

Bostock, D. (1979). *Logic and arithmetic* (Vol. 2). Oxford, England: Clarendon Press.

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

Campbell, N. R. (1920). *Physics, the elements.* London: Cambridge University Press.

Cliff, N. (1982). What is and isn't measurement. In G. Keren (Ed.), *Statistical and methodological issues in psychology and social sciences research* (pp. 3–38). Hillsdale, NJ: Erlbaum.

Dingle, H. (1950). A theory of measurement. *British Journal for the Philosophy of Science, 1,* 5–26.

Ellis, B. (1966). *Basic concepts of measurement.* London: Cambridge University Press.

Fechner, G. (1966). *Elements of psychophysics* (H. E. Adler, Trans.). New York: Holt, Rinehart & Winston. (Original work published 1860)

Field, H. (1980). *Science without numbers.* Oxford, England: Blackwell.

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

Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review, 91,* 139–150.
Heath, T. L. (1908). *The thirteen books of Euclid's elements* (Vol. II). Cambridge: University Press.
Helmholtz, H. von (1887). Numbering and measuring from an epistemological viewpoint. In P. Hertz, & M. Schlick (Eds.), *Hermann von Helmholtz: Epistemological writings* (pp. 72–114). Dordrecht, Holland: Reidel.
Hull, C. L. (1943). *Principles of behavior.* New York: Appleton-Century-Crofts.
Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement,* (2nd Ed.) Washington, DC: American Council on Education.
Krantz, D. H. (1972). A theory of magnitude estimation and cross modality matching. *Journal of Mathematical Psychology, 9,* 168–199.
Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of Measurement* (Vol. 1). New York: Academic Press.
Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751.
Roberts, F. S. (1979). *Measurement theory.* Reading, MA: Addison-Wesley.
Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16,* 170–233.
Russell, B. (1903). *The principles of mathematics.* London: Cambridge University Press.
Russell, B. (1983). The a priori in geometry. In K. Blackwell, A. Brink, N. Griffin, R. A. Rempel, & J. G. Slater (Eds.), *The collected papers of Bertrand Russell* (Vol. 1, pp. 289–304). London: George Allen & Unwin.
Spearman, C. (1927). *The abilities of man.* New York: Macmillan.
Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 667–680.
Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.
Stevens, S. S. (1959). Measurement, psychophysics and utility. In C. W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 18–63). New York: Wiley.
Suppes, P. (1951). A set of independent axioms for extensive quantities. *Portugaliae Mathematica, 10,* 163–172.
Suppes, P. (1959). Measurement, empirical meaningfulness and three-valued logic. In C. W. Churchman, & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (pp. 129–143). New York: Wiley.
Suppes, P., & Zinnes, J. L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 3–76). New York: Wiley.
Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34,* 273–286.
Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384–400.
Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.
Titchener, E. B. (1905). *Experimental psychology* (Vol. 1). New York: Macmillan.
Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin, 96,* 394–401.

Received September 16, 1985
Revision received March 12, 1986 ■