Measurement Scales and Statistics: The Misconception Misconceived

James T. Townsend
Purdue University

F. Gregory Ashby
Ohio State University

We reply to a recent *Psychological Bulletin* article by Gaito (1980) in which he assailed the proposition that fundamental measurement theory has any relevance to statistical theory and procedures. The major arguments Gaito adduced are evaluated and found to contain nothing that constitutes a logical or empirical refutation of the tenets of fundamental measurement theory or its implications for statistical analyses of data. Examples of statistical pitfalls to be met in ignorance of measurement considerations are given.

The implications of fundamental measurement theory for statistical analyses continue to be a disputed if not inflammatory issue. On one side is the view that the scale on which a set of measurements lies determines the type of statistical treatments that are suitable for application to the measurements. The opposing view is that there is no relation holding between the measurement scale and statistical procedures; essentially anything goes, relative to the measurement stipulations.[1]

Perhaps the most recent sally from the ranks of those adhering to the second position is that of Gaito (1980), who took the proponents of the measurement view to task. The potentially emotional nature of the issue can be appreciated from examples of invective in the pertinent literature, as in Gaito's labeling a contemporary measurement-oriented statistical book as reaching the "height of absurdity" (Gaito, 1980, p. 564). Such provocative rhetoric does little to clarify or decide the issue.

We defend what we refer to as the "measurement view" in this article. First, we outline this view, in a necessarily brief and qualitative manner.

Measurement Position

The basis of the measurement approach may be ascertained from standard sources (e.g., see Krantz, Luce, Suppes, & Tversky, 1971; Suppes & Zinnes, 1963; or Roberts, 1979, for rigorous definitions and developments). We remark that although Stevens (1951) was responsible for some critically important ideas regarding the use and misuse of measurement scales, the theory has progressed far beyond his work. Indeed, several of his most valuable concepts gain their true significance only in the context of the later developments. Yet many of the antagonists to the measurement view ignore the later foundational theory.

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.

There are two aspects of *fundamental measurement*. The first is the *representation theo-*

Requests for reprints should be sent to James T. Townsend, Department of Psychological Sciences, Purdue University, Peirce Hall, West Lafayette, Indiana 47907.

[1] As an aside, Gaito (1980) remarked that "there is no relationship between type of scale and statistical techniques used" (p. 564). Some of his other arguments are based on the presumed minor impact of measurement violations on statistical conclusions. It would seem that if the former argument were true, the latter would be impossible and vice versa.

Psychological Bulletin, 1984, Vol. 96, No. 2, 394-401

rem, which provides for the assignment of numbers to the empirical objects. Often, in the case of an infinite set of objects, one may simply prove that the assignment exists or state a method by which the numbers can be assigned.

Two scientists might by chance give two different assignments to a set of objects, each appropriately reflecting the critical empirical relational structure. In an important class of measurement systems, one may characterize the relation that has to hold between any two such numerical assignments as being in the form of a type of mathematical function mapping one of the assignments to the other. This illustrates the *uniqueness theorem,* which states the type or class of function that must relate the two numerical representations. Any transformation of an assignment (henceforth also called a *scale*) that is outside the permitted type distorts the numerical structure so that the original empirical relations are deformed or lost. Put another way, such an inappropriate transformation produces a wrong scale, one that is no longer a member of the set of allowable representations of the data.

It is generally agreed that the usual measurement of temperature through the expansion and contraction of mercury in Celsius or Fahrenheit units lies on an interval scale. It has been suggested that IQ might also lie on an interval scale. The designation as an interval scale implies that the unit of measurement as well as the origin (zero point) from which the measurement is taken may be altered without harm to the underlying physical or psychological qualities that one is attempting to capture. In contrast, if one were to square or take the logarithm of a temperature measurement, most scientists—in concert with the measurement theorists—would argue that the original real-world relations would be irrevocably distorted.

By the same token, statements about measurements of temperature can be either meaningful or meaningless.[2] To claim that the temperature today is twice as hot as it was yesterday may be true in terms of a specific pair of numbers, for example 64 °F versus 32 °F, yet without empirical content with reference to the physical reality. The reason is that the statement will no longer be true after a legitimate change of scale, to Celsius for example. In other words, there is no physical correspondent in the sense that the volume or molecular activity is doubled or the like. (Of course, if temperature is measured in degrees kelvin, then it lies on a ratio scale, and such statements are valid.)

If, in contrast, the difference in temperatures between July 4th and Christmas Day of 1982 is claimed to be twice the corresponding difference of 1981, the assertion is valid as a logically correct, but not necessarily empirically true, fact. It holds whatever (interval) scale is used to measure temperature.

Now one may object to the foregoing analysis, although it is difficult to envisage a cogent argument against it. However, if it (and by this we mean the rigorously developed theory) is accepted, then one is perforce led to a similar view of statistical manipulation and treatment. One may test mean group differences and perform an analysis of variance (ANOVA) on interval scale data because legitimate scale changes cancel out. Yet the same is not true of tests using the geometric rather than the arithmetic mean because such scale alterations do not cancel out and thus affect one's conclusions. Similarly, if the strength of one's data is only ordinal, as much of that in the social sciences seems to be, then even a comparison of group mean differences via the standard Z or t test or by analysis of variance is illegitimate. Only those statements and computations that are invariant under monotone (order is preserved) transformations are permissible.

Considerations of the Common Objections

We focus on the points Gaito (1980) made in collating typical views antagonistic to measurement implications for statistical procedures. As there is some overlap across positions, it is not necessary to cite all dissenters. We confess also that we were simply not able to make sense of some statements so here we concentrate on those that at least seemed sensical.

---

[2] As is perhaps obvious, meaningfulness is an all-or-none concept. Thus a statement can not be almost meaningful.

Fundamental Measurement Theory and Statistics are Mutually Independent

One approach Gaito adduced is to claim that measurement theory is valid within certain spheres but that its domain does not properly overlap with that of statistics.[2] In the statistical, as opposed to the deterministic, situation, the practicing scientist is confronted with a probability distribution on the measurement values of the objects treated. A probability distribution could arise because of the experimental or measurement error, or because of a more primitive variation in the qualities of the objects measured. In either event, in our opinion, the randomness of the quantities does not save them from the implications of the measurement theory.

In a variant of the independence assertion, it is claimed that the only pertinent constraints on the use of data are associated with the assumptions of the statistical model. The analysis of variance is advanced as an example here. The model assumptions are those of normality, equal variance, and the like (see any standard statistics text), and these do not generally refer to any concern with fundamental measurement. Our attitude is that the assumptions about the distribution type (e.g., normal), parameter values (e.g., equal variances), and the like, must be taken in addition to the constraints of measurement, in particular those provided by the uniqueness theorem. Why do most statistical texts fail to mention the measurement problem or teach the reader how to properly handle different scales statistically? The answer overlaps the next argument.

Statisticians Ignore or Attack Measurement View

In addition to the wholesale neglect of measurement questions by much of the statistical literature, some statisticians have spoken out against it (see Gaito for several references of this variety). None with which we are familiar refute the bases upon which fundamental measurement rests. Indeed, most statements seem beside the point, being simply derogatory, with little discernable content. For instance, Savage's (1957) remark that "I know of no reason to limit statistical procedures to those involving authentic operations consistent with the scale of observed quantities" (p. 340) is of this variety. But perhaps the best known, and certainly most often repeated statement in Gaito's articles is by Lord (1953): "The numbers do not know where they came from (p. 751)." Just exactly what this curious statement has to do with statistics or measurement eludes us. It is nevertheless worthwhile to examine the apparent basis for this remark.

The satirical article from which it was garnered (Lord, 1953) is based on the idea of generating data measured on a nominal scale and asking questions about whether samples are selected from the same distribution or not. The essence of the logic seems to be that it is just as legitimate to perform the usual t tests and such, on these samples as on any others; the specific appropriate probability distributions, and so on, are what count, not the so-called measurement scale.

Thus, in Lord's (1953) example, the freshman class is in a state of umbrage because the football jersey numbers assigned to them appear low relative to those given to students in higher grades. Suppose a significance test is carried out on the pertinent respective numbers and it is found that indeed, the numbers belonging to freshmen are significantly lower than those assigned to the other students at the .001 level. In fact, suppose that they were in fact drawn from a parent distribution with a lower mean. This is all fine and good. All the usual statistical theorems apply to these distributions (e.g., central limit theorem, Tchebyschev's inequality), subject to the following conditions. First, because the numbers were originally assigned only to uniquely identify the different players, the relative sizes of the football jersey numbers and the varying computations to which they are submitted do not imply anything about the real world. That is, the fact that the average freshman number is smaller than that of other students reveals nothing about the value of the football players wearing them, or anything else. Second, any one-to-one transformation may be visited on the numbers without changing any referential properties with regard to the real world, because they had none (except for naming) to begin with.

The outrage the freshmen suffered because of their miniscule numbers was an outcome of an invalid external interpretation, having absolutely nothing to do with the measurement scale or the way the numbers were assigned.

Perhaps a generous interpretation of the above contention would be to question whether a statistical analysis is ever of interest without reference to an empirical system. We believe not. With regard to the apparent general neglect by the statistical community of measurement considerations, there seem to be several reasons. One is simply that many of the quotations antedate published reports of the fundamental measurement theories. In particular, investigation of the representation problem was central to providing a firm underpinning of the uniqueness question and the consequent theory of admissible transformations. The latter results did not become available until the late 1950s and early 1960s.

Another factor may be that the most accepted of the axiomatic treatments of probability theory, providing the basis of much rigorous statistical work, is Kolmogorov's (1956) set-measurement theoretic foundational work. In its esthetic abstraction, there is no discussion of "real measured entities."

Receding further into the past, we find that some of the early and most influential statisticians worked either with ratio scales (and therefore naturally would have had little concern with the scale problem) or with populations and samples based on counting (e.g., number of people expiring between the ages of 60 and 65), thus inducing the even stronger absolute scales. Even today, when pressed to give a verbal, intuitive, or "real-world" explanation or description of probabilities, statistics, and asymptotic results, statisticians often rely on the *frequency explanation* of probabilities, again associated with counting, and therefore absolute scales. All in all, it may be that many statisticians have simply not taken the time to delve into these areas; it has probably not seemed to be a particularly important quest. We believe it is. Whether measurement questions will continue to be invisible in the statistical literature remains to be seen.

Statistical Robustness and Measurement Theory

At least one study has probed the effects of using illegitimate statistical techniques in data analysis. Baker, Hardyck, and Petrinovich (1966) transformed a set of random numbers using nonlinear, but order-preserving, functions and then performed standard t tests on samples from the original and the transformed data and compared the outcomes. Because conclusions from the two sets of tests generally agreed, it was argued that standard statistical techniques such as analysis of variance are robust with respect to the scale type of one's data.

Although this investigation is intriguing, it can logically reveal little or nothing about measurement scale robustness (with regard to statistical procedures). The reason is that in general we have no idea as to the degree of transformation that may occur in nature. The transformations used by Baker et al. (1966), while they are interesting and maintain monotonicity, fail to explore common monotonic transformations that stretch or shrink one part of the scale more than another, as does $f(x) = a \cdot x^2$, for instance. In the theory of the ordinal measurement, any monotonic transformation is admissible. Transformations that are near linear are not better or more admissible than are highly nonlinear transformations. They are not necessarily even more realistic. Without some knowledge about the kinds of transformations that may actually occur in nature, simulations such as those by Baker et al. (1966) cannot assess real-life robustness.

Suppose, for instance, that one were sufficiently untutored to attempt a test of the difference between two ratios of mean temperatures:

$$(T_{a1}/T_{b1}) - (T_{a2}/T_{b2}).$$

Let $T_{a1} = 64$, $T_{b1} = 32$, $T_{a2} = 100$, $T_{b2} = 50$ for Fahrenheit. Then the difference between the respective ratios is 0, but in Celsius it is infinite, or undefined. No amount of argument based on simulation or reasonable illegitimate transformations can save this operation from meaninglessness.

Similarly, suppose that peoples' racial prej-

udices lie on an ordinal scale. Then if two different scales are used, we are only guaranteed that order will be preserved in the responses. Thus, the probability distribution will be changed by the (unknown) monotonic transformation relating the scales. Will tests of significance between, for example, the means of rural and urban dwellers remain the same on the two scales? This is an empirical question that can be answered by actually performing the scaling and testing the two groups, but it cannot be answered by a priori simulation and transformation on artificial data. Furthermore, it is of the utmost importance to understand that even if the results of the test remain the same, the mean statistic and the test would still be meaningless: It is vacuous (or misinformative) with regard to revealing anything about the underlying structure of the phenomena of racial prejudice. Now let us turn to some detailed examples of what may be wrought by the willful misuse of statistics relative to laws of measurement.

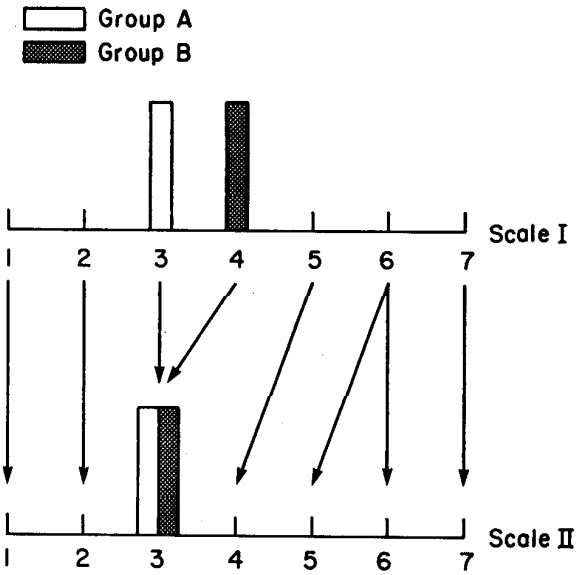

Figure 1. A transformation rendering a previously significant difference to a nonsignificant difference.

Examples

As a simple example of how a permissible transformation on an ordinal scale of measurement can change statistical conclusions about group differences, suppose two groups of subjects are asked to respond to a questionnaire on a 7-point (ordinal) rating scale. Suppose, as illustrated in Figure 1 (top), that all members of Group A respond 3 and all members of Group B respond 4. Certainly, any statistical test performed on these data will conclude that a real difference exists between the responses given by the two groups.

Now suppose the transformation illustrated in Figure 1 is performed. This transformation preserves the (weak) order of Scale I and so is an admissible transformation with ordinal scales. On the transformed Scale II, all members of both groups respond 3 and so any statistical test now concludes that no significant difference exists between the responses of the two groups. Clearly, in such a case, the statistical test is meaningless, that is, it tells nothing about the empirical system we are interested in.

As an example of how such a transformation might occur in an experimental setting, suppose there is some true underlying continuous scale on which all Group A members have a value of 3 and all Group B members have a value of 3.6. On Scale I, given a choice between 3 and 4, Group B members choose to respond 4. Now suppose that a second researcher independently administers the same questionnaire, with one exception, to an identical set of subjects. The exception is that instead of providing a response scale marked off with the numbers 1 through 7, the second research provides a set of seven ordered affective statements (strongly disagree, disagree, slightly disagree, and so on). For purposes of data analysis, assume this researcher later assigns these statements the numbers 1 through 7, respectively. It is possible that in this case, the same Group B population that responded 4 on Scale I will interpret the fourth affective statement of Scale II as having a value of, say, 4.5 on the true underlying scale and thus will choose the third statement as being most representative of its position. At the same time, Group A may still feel that the third response alternative accurately reflects its positions with the result that, as in Figure 1, all members of both groups have the same response.

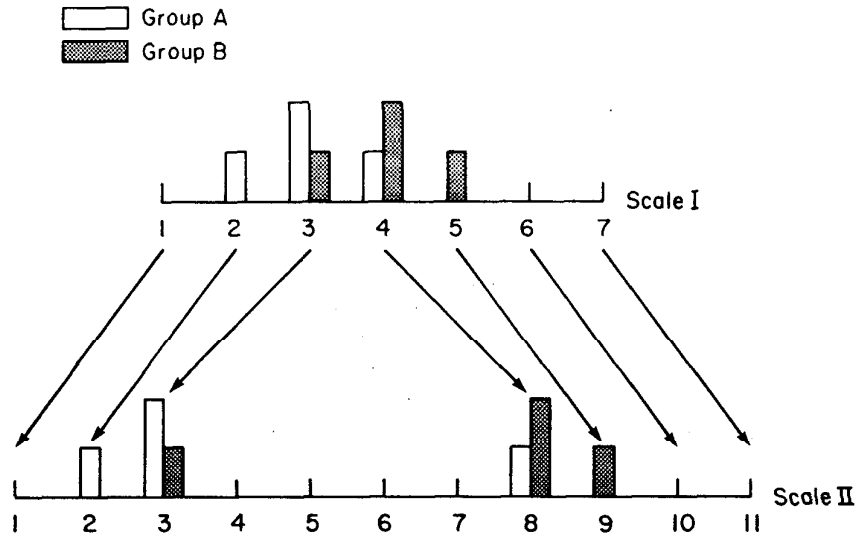

Figure 2. A transformation rendering a previously nonsignificant difference to a significant difference.

In this example, the empirical system under study is the same in both cases, as is the questionnaire. The only difference is in the particular ordinal scale chosen by the two researchers to measure the empirical attribute of interest. It is appropriate to use a *t*-test to decide that a significant numerical difference exists between the two groups on Scale I. However, it is inappropriate to conclude that a significant difference therefore exists within the empirical system. Thus, in this case, the finding of a significant statistical difference is a scale-dependent result because it is only true when the empirical attribute is measured on some of the permissible numerical scales. A statement about an empirical system is meaningful only when it is scale independent, that is, only when it is true on all of the permissible numerical scales.

If the scales of Figure 1 had been of interval level, the illustrated transformation would be inadmissible. Consequently, the statement that a significant empirical difference exists between the two groups would be meaningful, given the Scale I data of course.

It is important to realize that the point made by this simple example does not depend on the fact that responses 3 and 4 of Scale I are both mapped into the same response by the transformation illustrated in Figure 1. The only strictly increasing transformation from one 7-point scale to another is the identity (which maps response i to response i). Thus, a nonstrict monotonic transformation was used, but the conclusion does not depend on this feature. In fact, a similar example with a strictly increasing transformation is easily constructed if we allow the two scales to have a different number of response alternatives.

For example, consider the situation in Figure 2. On Scale I the responses of Groups A and B are bunched closely enough so that a statistical analysis is likely to conclude that no significant difference exists in the mean responses. After the strictly order-preserving transformation to Scale II, the responses of the two groups are largely separated, so that a statistical analysis would seem more likely to conclude that a significant difference exists between the mean responses of the two groups. Of course, if a z test or a t test is used to determine statistical significance, the variance must also be taken into consideration. The transformation relating the responses of the two groups increased the Group B mean relative to the Group A mean. However, at the same time the variance of each group's responses also increased and thus it is not unambiguously obvious that, say, the z score

$$z = \frac{\text{mean}_A - \text{mean}_B}{\sqrt{1/N(\text{variance}_A + \text{variance}_B)}},$$

where N equals sample size for each group, will be significant in the Scale II case but not in the Scale I case.

It turns out, though, that strictly increasing transformations with this property are fairly easy to construct. In fact, it is possible to find transformations with an even stronger property. We now present an example in which the Group A and Group B means are originally identical, but in which a strictly increasing transformation results in a highly significant z score.

Suppose Scale I is a continuous positive-valued ordinal scale and that the responses of Groups A and B are both log normally distributed on that scale with probability density functions

$$f_A(x) = \frac{1}{x\sqrt{2\pi(4\mu + \sigma^2)}} \exp\left[-\frac{(\log x + \mu)^2}{2(4\mu + \sigma^2)}\right]$$

and

$$f_B(x) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[-\frac{(\log x - \mu)^2}{2\sigma^2}\right];$$

$$\sigma > 0$$
$$0 \leq x < +\infty.$$

In this case the means of both distributions equal

$$E(X_A) = E(X_B) = \exp(\mu + \sigma^2/2),$$

and so the z score on Scale I of the difference between the group means is 0.

Now consider the strictly increasing transformation to Scale II, $Y = \log X$. In this case Scale II is continuous and defined on the entire real line (i.e., Scale II is both positive and negative valued). The Group A and B responses are now both normally distributed with means $E(Y_A) = -\mu$ and $E(Y_B) = \mu$ and variances $\text{var}(Y_A) = 4\mu + \sigma^2$ and $\text{var}(Y_B) = \sigma^2$. The z score for the difference between the group means on Scale II is therefore

$$z = \frac{2\mu}{\sqrt{4\mu + 2\sigma^2)/N}}.$$

It is easily seen that by selecting μ large relative to σ^2 results in an arbitrarily large z score. For example, if $N = 32$, $\mu = 7$, and $\sigma^2 = 2$, the resulting z score is about 14, a highly significant value. So here we have an example where a zero difference between groups has been changed to a significant difference.

Summary

Fundamental measurement theory asserts that operations and statements concerning measured quantities should not be performed if they violate those permitted by the uniqueness theorem that is pertinent to the measured quantities. The doubters attempt to confute this approach by various arguments, some of them self-contradictory, and the most important of these were considered. We find no logically and mathematically developed foundation for measurement and statistics that leads to a clear refutation of the fundamental measurement admonitions. Rather, most of the reasoning seems to be vague, qualitative, and occasionally emotional or defamatory.

A sensible question that has been raised (e.g., Baker et al., 1966) is whether the impermissible transformations that might occur by chance or due to one's choice of measuring instruments (and so on) are sufficient to alter one's statistical conclusions. Thus, if the true scale is interval, does a random change of the scale values lead to altered significance values? Unfortunately, without knowing the magnitude of such deformations in practice, such simulations offer little help. Nevertheless, it may be that future innovations will offer some solutions to this particular problem. What is worse, though, is that if one starts with an ordinal scale, comparing means (or some other illegitimate statistic) is simply meaningless, regardless of the magnitude of any further perturbations. We offered a few examples to illustrate how matters may go awry when the caveats of the measurement approach are not heeded.

It seems unlikely that uncompromising antagonists to the measurement approach will be convinced by the present reasoning, or any other for that matter. This article was prompted by an urge not so much to convert those individuals as to offer a countering view to the emerging generation of young behavioral scientists.

References

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement, 26,* 291–309.

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

Kolmogorov, A. (1956). *Foundations of probability* (N. Morrison, Trans.). New York: Chelsea. (Original work published 1933)

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, Vol. I.* 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.

Savage, I. R. (1957). Nonparametric statistics. *Journal of the American Statistical Association, 52,* 331–344.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). 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. 1–76). New York: Wiley.

Received April 4, 1983
Revision received January 24, 1984 ■