

# An empirical basis for the statement that measurement scale properties (and meaning) are irrelevant in statistical analyses

JOHN GAITO and RAY YOKUBYNAS  
York University, Downsview, Ontario, Canada

A study was conducted to evaluate the notion that the relationship between a number and its referent determines the type of statistical analysis required (the measurement-statistics issue). A number of transformations of original data were performed in which the meaningfulness of this relationship was modified. No change in statistical analyses resulted, even when meaningfulness was at zero, or near zero, levels with random transformations.

The controversy as to the independence, or the nonindependence, of measurement scale properties in a statistical analysis began with Stevens's (1946) thesis that the specific measurement scale involved with data (nominal, ordinal, interval, ratio) determines the specific operations of a statistical analysis. Recently, this thesis has been championed by Townsend and Ashby (1984) and rejected by Gaito (1980, 1986).

Gaito's (1986) argument is that, for measurement purposes, numbers are important because they relate to some underlying referent. However, in a statistical analysis, these referents do not enter the picture; it is only the numbers (which have no uniqueness except as numbers) that are involved in the statistical operations in a manner prescribed by the *mathematical properties* of the method. These statistical operations allow an effective ordering of the sets of numbers so that empirical statements (and associated meaning) can be added in interpretation of the results.

The purpose of this note is to provide an empirical basis for the statement that the meaning involved in the relationship between symbols (e.g., numbers) and their referents (although empirically important in measurement analyses) is irrelevant in a statistical analysis, because "The numbers do not know where they came from." That is, the numbers behave in the same fashion in a statistical analysis, no matter what their origin or scale properties (as Lord showed in a humorous article in 1953; also see Burke, 1953). The only aspect in the statistical analysis that is important is the relationship between the numbers (i.e., 1, 2, 3, 4, etc. are different from each other).

## METHOD

The experimental data was taken from an experiment by Toukmanian and Rennie (1975). Before and after a training session, two advanced

---

We thank Shake Toukmanian and Dave Rennie for providing us with the empathy data. The authors' mailing address is: Department of Psychology, York University, 4700 Keele St., Downsview M3J 1P3, Ontario, Canada.

doctoral students in clinical psychology rated the empathy of a number of counseling trainees (on a scale from 1 to 5). For the purpose of this study, only the pre- and posttraining ratings by one doctoral student were used (Session 1 and Session 2). The statistical analysis was a  $2 \times 12$  (sessions  $\times$  subjects) design, using the original data and a number of transformations: reversing the scale, adding and subtracting a constant of 2, multiplying and dividing by a constant of 2, and randomizing the set of numbers separately within the pre- and posttraining sessions. Two separate randomizations were used,  $R_1$  and  $R_2$ .

The various transformations of the empathy scale provided for a variation of meaning. The scale from 1 to 5 is the most meaningful, because a low number of 1 means an *absence of empathy*, and each succeeding and adjacent number increases the level of empathy. Furthermore, the 1-to-5 set is used frequently in scales, thus indicating a consensual expression of meaningfulness. The constant transformations also are meaningful in that the relative position of each number to the others remains the same, but the relationship between a number and the referent appears to be less meaningful because different numbers are associated with the same categorization or referent. For example, a 1 is associated with *absence of empathy* in the original data, but performing the constant transformations provides the numbers 3, -1, 2, and .5 for this referent. Obviously, 1 and .5 appear more logical and meaningful than 3, -1, and 2 for *absence of empathy*. A similar statement can be made for the reversal of the scale. The random transformations provide for zero (or low) degrees of meaningfulness in that the relationship between the number and the referent is a random, non-systematic one.

## RESULTS

The statistical results using ANOVA were the same for the reversed condition, the conditions using constant transformations, and the original condition. For each of these analyses,  $F(1,11) = 10.70$ ,  $p < .05$ , indicating differences between the two sessions in mean ratings of empathy, with the posttraining mean being the greatest. Similar results, and the same statistical conclusions, occurred with the two random conditions. For  $R_1$ ,  $F = 8.09$ ; for  $R_2$ ,  $F = 9.48$ . These slight changes in  $F$  values appeared because with random assignment of numbers in  $R_1$  and  $R_2$ , the SS for sessions remained the same, but the subjects' SS changed slightly.

Identical statistical conclusions would be obtained if the design had been an independent-subjects ANOVA. For example, if the pre- and posttraining empathy ratings were

for different groups of subjects, such as interns and experienced clinicians, then  $F(1,22) = 11.58$ ,  $p < .05$ , would result for any random ordering of numbers within each particular group.

Table 1 shows scale values for various conditions and for two subjects. Subjects 7 and 11 in the original condition provide the standard (real and most meaningful) values for comparison with those from the other conditions. The numbers with the four constant transformations, the reversed condition, and the random condition provide different scale values in each case. For example, Subject 7 for Session 1 has a 2, indicating a low amount of empathy for the original condition, but the scale values for this same degree of empathy become 4, 0, and 1 for the constant transformations, and 4 for the reversed condition.

The greatest discrepancy relative to the original condition (and low or zero meaningfulness) occurs with  $R_1$  and  $R_2$ . In the original condition, Subject 7 shows no change in empathy rating from Session 1 to Session 2, but an increase in empathy rating occurs in  $R_1$  and  $R_2$  (1 to 2.5 and 1 to 3, respectively). On the other hand, Subject 11 has a marked change in empathy rating from Session 1 to Session 2 in the original condition (1 to 3), a regression in  $R_1$  (2 to 1), and a slight increase in  $R_2$  (2 to 2.5).

Yet in all of these conditions, the statistical results are similar. Thus, although meaningfulness varies from one condition to the next (in terms of the relation between the number and its referent in the original condition), and reaches a low or zero point in the random conditions, the statistical conclusions are the same. These results show clearly that scale aspects are irrelevant for statistical analyses. The contrast between the original condition and random condition for Subjects 7 and 11 is most pronounced: the range of scale values stay the same; the statistical conclusions overall stay the same; but Subjects 7 and 11 obtain different scale numbers, indicating different degrees

of rated empathy than each individual had in the original condition.

## DISCUSSION

The results of the present empirical study add further weight to the statement that measurement scales are irrelevant in a statistical analysis. Although the number and the referent (and meaningfulness) varied, the statistical conclusions concerning mean differences between the groups of data remained the same. The striking result was that the same conclusions remained even in the condition in which numbers were randomly assigned to each subject within each group, producing drastic differences in degrees of empathy which were not present. Thus, meaningfulness here would be at a low or zero level. The discrepancy between the original condition and the random ones would have been greater if all scale values (1-5) had been obtained in Sessions 1 and 2.

There are a number of other points that can counter Stevens's (1946) argument and that complement these empirical results (Gaito, 1980, 1986). These include the fact that data can consist of a single response of a subject, a number of responses by one subject, or a mean of the number of responses for one or more subjects. These possible data would consist of different scales. Furthermore, different statistical procedures listed as involving data from different scales provide the same statistical result (e.g., binomial test and sign test; binomial distribution and normal approximation; and  $\chi^2$ , binomial distribution, and normal approximation when 1 *df* is present).

The results of the present study are consistent with the thoughts of Burke (1953), Lord (1953), Anderson (1961), Gaito (1960, 1980, 1985), Binder (1984), and others, that measurement scale values are irrelevant in a statistical analysis. To put it simply, "The numbers do not know where they came from."

## REFERENCES

- ANDERSON, N. H. (1961). Scales and statistics: Parametric and non-parametric. *Psychological Bulletin*, **58**, 305-316.
- BINDER, A. (1984). Restrictions on statistics imposed by method of measurement: Some reality, much mythology. *Journal of Criminal Justice*, **12**, 467-481.
- BURKE, C. J. (1953). Additive scales and statistics. *Psychological Bulletin*, **60**, 73-75.
- GAITO, J. (1960). Scale classification and statistics. *Psychological Review*, **67**, 277-278.
- GAITO, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, **87**, 564-567.
- GAITO, J. (1986). Some issues in the measurement-statistics controversy. *Canadian Psychology*, **27**, 63-68.
- LORD, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, **8**, 750-751.
- STEVENS, S. S. (1946). On the theory of scales of measurement. *Science*, **103**, 677-680.
- TOWNSEND, J. T., & ASHBY, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, **96**, 394-401.
- TOUKMANIAN, S. G., & RENNIE, D. L. (1975). Microcounseling versus human relations training: Relative effectiveness with undergraduate trainees. *Journal of Counseling Psychology*, **22**, 245-352.

(Manuscript received for publication July 19, 1986.)

Table 1  
Variation in Scale Values for Two Subjects (7 and 11)  
Over a Number of Conditions

Condition	Possible Scale Values	Actual Values		$S_7$		$S_{11}$	
		Pre	Post	Pre	Post	Pre	Post
Original	1 to 5	1,2	1 to 3	2	2	1	3
Revised	5 to 1	5,4	5 to 3	4	4	5	3
+2	3 to 7	3,4	3 to 5	4	4	3	5
-2	-1 to 3	-1,0	-1 to 1	0	0	-1	1
$\times 2$	2 to 10	2,4	2 to 6	4	4	2	6
$\div 2$	.5 to 2.5	.5,1	.5 to 1.5	1	1	.5	1.5
Random 1	1 to 5	1,2	1 to 3	1	2.5	2	1
Random 2	1 to 5	1,2	1 to 3	1	3	2	2.5