# The Discriminating Power of Ordinal Data

David J. Weiss
*Department of Psychology*
*California State University, Los Angeles, CA 90032*

*Ordinal information was extracted from temperature readings from Los Angeles and Minneapolis. The analytical power of several ordinal transformations was compared to that of the raw data. Rank orders were as effective as raw temperatures in terms of discriminating between months, and this equality of effectiveness did not depend on how many scores per month were analyzed. A rating scale employing 20 categories was virtually as powerful also; scales with 6 and 2 categories were respectively less effective. It was suggested that Stevens' (1946) hierarchy of scale types is not a useful system of classification, and that ordinal scales should not automatically be considered poor instruments.*

Suppose we were to measure temperature with an uncalibrated thermometer, one which would allow accurate ordinal comparisons but which would furnish no numerical values. With such a crude instrument, would it be possible to make correct decisions involving temperature, such as whether one room was reliably colder than another, or whether various time periods experienced equally warm temperatures? This question is of interest to a student of behavior because many behavioral measures are akin to the uncalibrated thermometer, in that comparisons of rank are meaningful but questions of amount are not. For example, I can assert confidently that I am hungrier now than I was after breakfast, and also that my hunger has increased since I could smell dinner cooking. However, I would be hard pressed to scale these subjective experiences with numbers that accurately conveyed their intensity. It is clearly easier to accept the face validity of ordinal comparisons than to believe that subjective ratings reflect internal experience in a numerically accurate sense.

Since the classic work of Stevens (1946), though, measurement theorists have stressed the importance of obtaining measures which have at least interval properties. The primary technical reason for this stress is no longer generally considered correct (Prytulak, 1975),

Author's Note: *I wish to thank Robyn Dawes and James Ramsay for helpful comments on an earlier version of this manuscript.*

*Journal of Social Behavior and Personality,* 1986, Vol. 1, No. 3, 381-389.

although the controversy has not died completely (Gaito, 1980; Townsend and Ashby, 1984). Stevens argued that interval measures are required for the customary statistical tests, such as ANOVA, because computing means and variances requires information about distances between points. Elegant papers by Anderson (1961) and Lord (1952) have exposed the flaw in this logic, that numbers in an analysis do not know their origin. Certainly the results of a statistical test are only as meaningful as the scores on which they are based, but the numbers themselves may be compared. Lord's example of a comparison of two sets of football numbers was especially forceful; one could compare the jersey numbers sensibly, although they would shed no light on any attribute relating to football.[1] While such comparisons may not be useful in a substantive context (Maxwell and Delaney, 1985; Prokasy, 1961), it is not the fault of the response scale; the obvious problem is the validity of the measure.

Still, the obvious advantage of interval measures remains; they contain more information than ordinal measures. However, this advantage may be more apparent than real. The pioneering efforts of Shepard (1966) showed the feasibility of extracting meaningful interval information from ordinal data. Shepard argued that the network of interrelations among the interpoint distances constrains ordinal values successively more tightly as the number of data points increases, so that eventually the information available in a set of ordinal data approximates that available in a corresponding set of interval data.

While Shepard's argument is intuitively plausible, it is difficult to understand his demonstrations because the recovery of metric structures requires a complex computer program (Shepard, 1962) to transform the data. The feeling that computer magic is necessary to retrieve metric information from interval data is also present when one examines later demonstrations, such as that of Weiss and Anderson (1972), which used a program written by Kruskal (1965). The present inquiry explores the relation between interval and ordinal data in a much simpler way.

Instead of looking at transformed ordinal data and comparing the derived scores with "true" values, we use the ordinal data to make substantive inferences using analysis of variance. The F-ratios are then compared with F-ratios based on "true" interval data. If the F-ratios are not significantly different, then the ordinal data are as accurate as the interval data. The comparison of F-ratios may be carried out with the test proposed by Bradley and Schumann (1957).

[1] *In the years since Lord selected his example, organized football has adopted a coding which associates the player's number with his position. Position is in turn roughly associated with football related attributes such as weight and speed. Currently, basketball numbers would furnish a more apt illustration.*

Because the F-ratio is a measure of the sensitivity of the response scale (Schumann & Bradley, 1957), it affords a convenient means of parametric exploration. Is the number of points in the data set crucial, as Shepard has suggested? Does the range of values being measured matter? One might expect that ordinal measures would have a tough time capturing the interpoint distances in a data set in which the true scores are tightly packed. And what about the precision of the ordinal scale? One would certainly expect a fine-grained scale, such as a pure rank ordering, to yield more information than a coarse scale with a few, relatively wide, categories.

## BASIC PROCEDURE

The basic data for the study were the daily high temperatures during 1983 for Minneapolis and Los Angeles. These data sets were extracted from U.S. Dept. of Commerce figures by John O'Hagan of the CSULA Dept. of Geography, to whom I am extremely grateful. Weather data were used, rather than behavioral data, because their character affords unquestionable face validity to substantive inferences. The scores were grouped by months, separately for each city, and 1-way ANOVA (with 12 groups) was performed on each set. Thus the trivial hypothesis that the various months are equally warm was the foundation for evaluation. This approach assumes (surely incorrectly) that all of the temperatures within a month have the same true value, and differences among them merely represent random error. The more incorrect this assumption is, the lower will be the power of the ANOVA in terms of detecting differences among the months. The huge F-ratios observed suggest that lack of power is not an important concern.

Then ordinal information was extracted from the data sets, and the resulting values were subjected to the same 1-way ANOVA procedure. The transformations were chosen to preserve the ordinal relationships among the scores while surrendering metric relations. The rank order transformation simulated comparative judgments, while the categorizing transformation simulated the ordered classification of rating procedures. The primary question of interest is whether the transformed responses, which have only an ordinal relationship to the underlying variable of temperature, will be as sensitive to the differences between months as are the raw scores. This sensitivity is given by the F-ratio, so it is the F-ratio which is the dependent variable in the present analyses. Of secondary interest is whether such parameters of the data set as the number of scores, or their density, affect the correspondence between results for raw and transformed data.

Table 1 F-Ratios

| | *Minneapolis* | *Los Angeles* |
|---|---|---|
| Full Month | Raw: 203.3 | Raw: 39.3 |
| | Rank: 219.2 | Rank: 42.3 |
| 15 Scores per Month Removed | Raw: 110.0 | Raw: 17.5 |
| | Rank: 121.8 | Rank: 17.7 |
| 25 Scores per Month Removed | Raw: 41.5 | Raw: 7.6 |
| | Rank: 39.6 | Rank: 9.4 |
| | | *F-Ratios for Category Scales* |
| 20 Categories | 193.9 | 40.3 |
| 6 Categories | 160.0 | 33.2 |
| 2 Categories | 87.2 | 30.2 |

## SPECIFIC PROCEDURES AND RESULTS

The weather stations were chosen to represent extremes of homogeneity of temperature, and the F-ratios using the raw temperatures confirm the selection. The range of temperatures in Minneapolis was from -16 to 97 (degrees Fahrenheit), while Los Angeles temperatures ranged between 40 (an aberrant day in April; the second coldest high temperature was 54) and 103. The F-ratios for the two cities were 203.3 and 39.3 respectively.

The first ordinal conversion was simply to replace each score with its rank within the data set; tied scores were given equal ranks. The ANOVAS on these ranks yielded F-ratios which were, somewhat surprisingly, slightly *higher* than those for the raw temperatures; they were 219.2 and 42.3 for Minneapolis and Los Angeles. While unexpected, this increase in the F-ratio can perhaps be understood by realizing that the ranks in each data set span the range between 1-365, a larger range than the temperatures which form the original data.

Our first conclusion is at hand. Apparently the range, or density, of the raw data does not play a role in whether ranks yield discriminability equivalent to raw data.

### Size of the Data Set

The number of scores per month was drastically reduced by randomly eliminating first 15, and then 25, of the temperatures in each month. This did not affect the degrees of freedom in the numerator of the F-tests, but the degrees of freedom in the denominator were decreased considerably. The reduction in power would naturally decrease the F-ratios, but would the rank orderings maintain their ability to discriminate among months as well as raw temperatures?

The F-ratios in Table 1 show that rank orderings (which no longer extend from 1-365, but now cover 1-174 and 1-53 for the two reductions) do not yield less discriminating power than the raw scores on which they are based, even for fairly small data sets. The Los Angeles, 25 scores deleted/month, data highlight the fact that even when the effect being measured is relatively small, rank orders do as good a job in extracting it as raw scores.

### Category Scales

Complete rankings are rare in research. Most ordinal data sets consist of category ratings. Rating scales were simulated by partitioning the range of temperatures for each city into *k* equally wide categories, then replacing each raw score with its category. The number of categories, *k*, investigated was either 20, or 6, or 2. Using two categories is like employing a "yes-no" response; here it might be though of as a "cold-hot" response.

The F-ratios for category scales in Table 1 contain no major surprises. As one would expect, the fewer the categories, the smaller the F-ratio. However, the rate of decrease in the F-ratio as the number of categories drops is quite low, especially for the dense Los Angeles data. The only significant (at the .05 level, using a 1-tailed Schumann and Bradley test) differences among these F-ratios and also the raw score F-ratios, comparing within-city, are that the 2-category F-ratio for Minneapolis is significantly smaller than any of the other Minneapolis F-ratios.

The conclusion to be drawn here is at odds with our first conclusion. Here, the range of the data does matter. Using a small number of categories loses information when the data cover a wide range, but produces less of a decrement when the data cover a narrower range.

## DISCUSSION

The present results suggest that the classification of scale types championed by Stevens (1946), an idea which is presented in many elementary statistics texts (e.g. McCall, 1980), may not be pertinent to substantive inferences. A continuum of precision in response systems seems a more useful concept. One may think of pure ranking as the employment of a category scale in which the number of categories is equal to the number of objects measured. The present investigation shows a steady decrease in the accuracy with which an ordinal scale captures the true differences between months as the number of categories decreases.

Peculiarly, ranks appear to be even better than the raw interval data; at least, higher F-ratios are produced. Is this then evidence that the F-ratio criterion is inappropriate? How can the ranks yield discrimination which is superior to the original data? Mathematically, the higher F-ratio for ranks is not startling; any nonlinear transformation changes the F-ratio, and one cannot predict the specific nature of that change. But is the increased F-ratio mere artifact? In the case of the Los Angeles data, it may be argued that the assessment of temperature based on ranks is indeed more accurate than that based on Fahrenheit temperature. Using the raw data, April was found to be the coldest month. That single aberrant cold day reduced the mean temperature so much that an "incorrect" average resulted. Using ranks, though, December was the coldest month. Averaging the ranks led to a conclusion that Los Angeles residents would find sensible, but averaging the raw temperatures led to a conclusion that lacked face validity. Thus it would appear that using rank order data rather than interval data may lead to a more appropriate analysis of variance in just the same way (and for the same reason) that a median may afford a more appropriate typical value than a mean.

The superiority of ranks has been studied in a different context by Iman and his colleagues (Iman, 1974; Conover and Iman, 1981), who have been interested in statistical power. They have conducted simulations employing the rank transformation. Data have been replaced by their ranks and subjected to the same analyses as the original data. This replacement procedure is identical to the basic scheme followed here. The general conclusion from this statistical literature is that analyses on ranks are almost never less powerful than those on raw data. For some underlying distributions (e.g., contaminated normal, which is the result of mixing two widely separated normal populations), they are considerably more powerful. This conclusion may reflect the fact that the sampling distribution of the t-statistic, and therefore the F-statistic as well, is not much affected by order-preserving nonlinear transformations when sample sizes are roughly equal (Baker, Hardyck, and Petrinovich, 1966).

The rank order question has also been considered analytically by Abelson and Tukey (1963), who showed that numerical values could be assigned to objects measured ordinally when extra information is available but not well specified. They stressed that in most applications, the researcher's perspective on an ordinal response instrument is that more than rank orders are known. Typically the excess knowledge is that the scale is reasonably smooth.

There is one sense in which the simulation of judgmental processes employed here using temperatures may be inaccurate. In converting raw temperatures to the various ordinal scales tested, no errors were made. Human judges engaged in rating are doubtless subject to random fluctuations. However, we need not be too concerned with the random errors in ranking which humans would make, because it is easy to predict their effects. The more misclassification, the lower the F-ratio. Obviously, the less precise is the ordinal scale (that is, the fewer categories), the more serious will be the effect of a given level of inaccuracy.

The present conclusion regarding the number of response categories, that more is better, is consistent with that of Garner (1960). Garner employed an information transmission criterion to evaluate rating scales. His criterion measured the extent to which different stimuli lead to different responses, and is closely related to the present F-ratio criterion (Garner and McGill, 1956). Garner argued that the optimal number of response categories depends on the discriminability of the stimuli being judged. If all of the objects are clearly discriminable, then there should be as many categories as there are objects so that the judges may display their rating prowess. In the case of complete discriminability, ranking will be the most effective procedure.

The practical conclusion of the present study is a simple one. Those who employ ordinal measures in substantive investigations need not feel apologetic. So long as the judge can order the stimuli, the technical aspects of the response system are not likely to be crucial to the drawing of correct inferences. As a general rule, the more categories, the better. Complete ranking, if possible, is at least as accurate as an interval scale. The numbers on the thermometer are not needed to compare the coldness of the various months. An empirical example of the effectiveness of ordinary rating scales has been furnished by Dawes (1977), who showed that members of a department can judge their colleagues' heights accurately using verbal ratings. The resulting estimates were highly consistent with physical heights. The present results should not be interpreted to mean that researchers should replace actual data with ranks. Numerical data obviously carry information beyond that to which analysis of variance responds. The main point is that even if the data are only of an ordinal nature, correct comparisons among groups can be obtained.

The important property of a response instrument is its ability to allow the stimuli to be ordered correctly. From the substantive perspective, the extent to which the responses have the more advanced properties of ordinary numbers, such as the equal interval or equal ratio properties, is not critical. This is probably fortunate, for in applied settings it is likely to be quite difficult to establish such properties (Wolins, 1978). The researcher who is in doubt as to the classification of a scale does not have to settle for a perhaps less powerful nonparametric test (Anderson, 1961). Ordinary analysis of variance will assess group differences correctly.

As Cohen (1968) has so cogently demonstrated, analysis of variance and regresion may be viewed as two sides of the same coin. One might speculate that comparisons of regression-based statistics, such as sets of correlation coefficients, would also be insensitive to whether the data were collected with interval scales.

## REFERENCES

Abelson, R.P., & Tukey, J.W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics, 34,* 1347-1369.

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

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

Bradley, R.A., & Schumann, D.E.W. (1957). The comparison of the sensitivities of similar experiments: Applications. *Biometrics, 13,* 496-510.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70,* 426-443.

Conover, W.J. & Iman, W.J. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *American Statistician, 35,* 124-129.

Dawes, R.M. (1977). Suppose we measured height with rating scales instead of rulers. *Applied Psychological Measurement, 1,* 267-273.

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

Garner, W.R., (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*, 343-352.

Garner, W.R., & McGill, W.J., (1956). The relation between information and variance analysis. *Psychometrika, 21*, 219-228.

Iman, R.L. (1974). A power study of a rank transform for the two-way classification model when interaction may be present. *Canadian Journal of Statistics, 2*, 227-239.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society* (Series B), *27*, 251-263.

Lord, F.M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8*, 750-751.

Maxwell, S.E. & Delaney, H.D. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin, 97*, 85-93.

McCall, R.B. (1980). *Fundamental statistics for psychology* (3rd ed.) New York: Harcourt Brace Jovanovich.

Prokasy, W.F. (1961). Inference from analysis of variance of ordinal data. *Psychological Reports, 10*, 35-39.

Prytulak, L.S. (1975). Critique of S.S. Stevens' theory of measurement scale classification. *Perceptual and Motor Skills, 41*, 3-28.

Schumann, D.E.W., & Bradley, R.A. (1957). The comparison of the sensitivities of similar experiments: Theory. *The Annals of Mathematical Statistics, 28*, 902-920.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*, 125-140.

Shepard, R.N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3*, 287-315.

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.

Weiss, D.J., & Anderson, N.H. (1972). Use of rank order data in functional measurement. *Psychological Bulletin, 78*, 64-69.

Wolins, L. (1978). Interval measurement: Physics, psychophysics, and metaphysics. *Educational and Psychological Measurement, 38*, 1-9.