COMMENTARY

Treating Ordinal Scales as Interval Scales: An Attempt To Resolve the Controversy

THOMAS R. KNAPP

Ever since the Harvard psychologist S. S. Stevens *(1946)* advanced his ideas concerning the connections between measurement scales and statistical analysis, there has been a continuing controversy regarding the use of traditional descriptive and inferential statistics for ordinal-level variables. The exchange in *Nursing Research* between Armstrong and this writer (Armstrong, *1981;* Knapp, *1984,* Armstrong, *1984)* is illustrative of the competing positions. In this commentary, an attempt will be made to sort out the dimensions of the controversy summarize some recent contributions that help to clarify the issues, and suggest a possible solution to the problem. The reader who is interested in the earlier history of conflicting views about scales and statistics is referred to the excellent and balanced review by Gardner *(1975).*

Psychometric Considerations

The first controversial matter is the determination of the type of scale one actually has Stevens (19-16) clearly demonstrated the difference between ordinal and interval scales. An ordinal variable arises from a scale for which all order-reserving (monotonic) transformations are admissible; that is, they leave the scale form invariant. For an interval variable, the only admissible transformations are those of the linear-form y = bx + a. (Transformation from Centigrade to Fahrenheit, by F = 1.8C + 32, is the classic example.) For a scale to be at the ordinal level of measurement, the categories comprising the scale must be mutually exclusive and ordered. The additional specifications that are typically postulated for an interval scale are the existence of an arbitrary zero point and an arbitrary unit of measurement that is constant throughout the scale. But pinning down scale type is not easy.

First, there are no agreed-upon rules for determining whether a particular scale is ordinal, less than ordinal, or more than ordinal. Consider, for example, the scale consisting of the following categories: never, seldom, frequently, always. Most researchers would agree that those four categories are mutually exclusive and ordered, in the sequence provided. Assigning the numerals *1, 2, 3, 4* to the four categories seems to be a reasonable thing to do; but any order preserving transformation (e.g., *1, 3, 5, 7* [linear] or *3, 8, 15. 16* [non-linear]) would seem to be equally admissible, and most would agree that such a scale is indeed ordinal. But suppose that the ordinal categories were: never, occasionally, sometimes, always. There would be little agreement among judges (and grammarians) about the relative placement of the two middle categories, thus producing a less-than-ordinal scale.

Second, does the shape of the distribution of scores on the variable have any relevance for its ordinality or intervality? Yes, argued Gaito *(1980),* Borgatta and Bohrnstedt *(1980),* and others who claimed if a variable is normally distributed it must constitute an interval scare. But Thomas *(1982)* disagreed; he proved the position "If normality; .then an interval scale" to be false.

The psychometric aspect received considerable attention in the early literature and in very recent years. Two of the latest contributions are Michell *(1986)* and Marcus-Roberts and Roberts *(1987).* Michell claimed that much of the disagreement regarding scale type (and the associated controversy about appropriate statistics) is attributable to differences in theoretical perspectives . on measurement. Scientists espousing the representational approach that Stevens assumed make certain distinctions which adherents to the operational or the classical theories find to be irrelevant. The operatonalists (among whom are counted most psychologists) are interested only in consistent assignment of numbers to objects in order to study quantitative relationships -between manifest variables and/or their underlying latent counterparts. Classical measurement theory deals only with the assessment of quantity, and all variables are of the same scaie t\ pe. Neither the operational nor the classical view shares the representational goal of drawing scale-free conclusions from scale-specific statements.

Marcus- Roberts and Roberts *(1987)* considered both psychometric and statistical matters. Their notions of "meaningful statements" and "meaningless statistics" provide the bridge between the psychometric issues in the ordinal /interval controversy and the issues involving descriptive and inferential statistics.

Descriptive Statistical Considerations

Most of the conflict between the pro-Stevens ("conservative") and the anti-Stevens ("liberal") camps begins after both sides agree that a certain variable is ordinal. But they part company when analyzing the data generated by that variable.

The "liberals' , argue that although they do not have a_ true interval scale, they regard the differences between categories *A* and B, B and C, etc. as equal. They also appeal to the work of Baker, Hardyk, and Petrinovich (1966), Labovitz (1967), and others who have shown empirically that it matters little if an ordinal scale is treated as an interval scale. The conservatives" counter that researchers aye demonstrated very strange results when using means, standard deviations, .and Pearson r's .with ordinal scales (Stevens, 1955; Otrien, 1979; Marcus-Roberts & Roberts, 1987). The mean for Group I can be higher than the mean for Group II on the original scale but lower on an admissible transformation of the original scale. The Pearson r for two ordinal variables can actually be opposite in sign to.. Spearman`s rho for t those two variables.

The keys to this dimension of the controversy are the notions of "appropriateness" and "meaningfulness." That is, what descriptive statistics are appropriate for ordinal scales? What statements regarding data reduction are meaningful?

Marcus-Roberts and Roberts (1987) argue that it is always appropriate to calculate means (for example) for ordinal scales,-but that it is not appropriate to make certain statements about such means. Drawing upon the original work of Suppes (1959) and others on the concept of meaningfulness, they show that a statement of :he form "the. mean for Group I on Variable X is greater, than the mean for Group II on Variable X" is meaningful for interval scales and meaningfulness or ordinal scales. The reason is that the claim always holds for admissible (linear) transformations of interval scales and does not hold for some admissible (order-preserving only) transformations of ordinal scales. Furthermore, meaningfulness is not the same as truth. They give as an example the statement: "I am twice as tall as the Sears Tower." That statement is meaningful, since its truth or falsity is invariant under any admissible transformation, but is obviously false for all scales that measure height.

Inferential Statistical Considerations

The arguments in the ordinal/interval controversy that relate to statistical inference are the hardest to untangle. The conservatives seem to believe that once, you reassigned to the ordinal level of description, you, I, are automatically restricted to inferences regarding population medians and modes rather than means, and rank correlations (or similar indices) rather than Pearson r's you must use non-parametric procedures rather than parametric procedures; and you are destined to have lower power. (See, for example, Siegel, 1956, p. 20.) The liberals see nothing wrong with means and Pearson r's for their ordinal scales they use the same parametric procedures for ordinal scales that they use-for interval scales, claiming that scale type is not included among the assumptions for the validity of the t and- F sampling distributions; and they insist, that they always have greater power. (See, for example, Labovitz, 1967, p. 158.)

Both sides have flaws in their arguments. As previously mentioned, Thomas (1982) showed that one can have a normal distribution for an ordinal scale. If such. a distribution is the population distribution for which the mean median and., mode are all the same, one should-use the sample mean to estimate the population mean or to test hypotheses about it, ordinal scale and meaningfulness to the contrary notwithstanding. The liberals are right in that interval level or above is not and never has been one of the general linear model assumptions. (Siegel 119561 was simply in error on this point and was properly taken to task by Armstrong [1981) and others.) But both camps are mistaken regarding.. the, alleged power superiority of parametric tests over non arametric tests. The wilcoxon tests for independent samples and for paired samples are never much less powerful than t, and when the population distribution is not normal (for ordinal or interval measurement) they can be much more powerful (Blair & Higgins, 1980; 1985).-The confusion here is apparently between robustness. and power. Appeals to the robustness of the t and F_ sampling distributions (Havelicek & Peterson, 1974) served to demonstrate that one can usually tease normality and homogeneity- of variance quite a bit without doing serious injustice to t or F, particularly with equal sample sizes (but see Trachtman, Ciambalvo, and Dippner, 1978). It doesn't necessarily follow, however, that you will always attain greater power in the process.

Recent work by Maxwell and Delaney (1985) on what happens to construct validity when one chooses over Wilcoxon, and by Marcus-Roberts and Roberts (1987) on tests of meaningful hypotheses, should also help considerably in clarifying the inferential statistical issues in the scales controversy. Maxwell and Delaney show that an independent samples t test for an observed variable produces the correct inference regarding the underlying construct only under very special conditions. Population medians involving order preserving transformations are not subject to such constraints.

Marcus-Roberts and Roberts (1987) argue, as they do for descriptive statistics, that though it may be appropriate to test nonmeaningful hypotheses such as $\mu_1 - \mu_2 = 0$ for an ordinal scale, it is meaningful hypotheses that should be tested. They define a meaningful hypothesis as a hypothesis H which holds if and only if T(H) holds, where T is the set of all admis sible transformations of the scale that are employed.

An Attempt to Resolve The Controversy

Although this writer is not naive enough to believe that he can end the war by methodological Bat, in the spirit of Adams, Fagot, and Robinson (1965)-a largely ignored theoretical contribution which actually tried to placate both sides over twenty years ago-he would like to recommend the following plan to researchers who are concerned about measurement scales and the corresponding statistical analyses:

1. Choose a measurement perspective (Michell, 1986). The representational theory of Stevens and others is fine if its goal of scale-free conclusions is commensurate with your philosophy of science. If it is not, and if you are attracted to the operational or classical theories, much or all of the scales-and-statistics controversy in the literature (and in this paper) is then irrelevant.

2. Read Stevens's (1996) first paper carefully. It is very well written and makes a lot of sense, especially in

the context of recent developments regarding the concept of meaningfulness. At one point in that paper, he even makes a practical concession to the "liberals" when he says:

> In the strictest propriety the ordinary statistics involving means and standard deviations ought not. to be used with these (ordinal] scales, for these statistics imply a knowledge of something more than the relative rank-order of data. On the other hand, for this 'illegal' statisticizing there can be invoked a kind of pragmatic sanction: In numerous instances it leads to fruitful results.

3. When you're making- up your scale (operationalizing your construct, choosing your categories, etc.), honestly face up to Stevens's taxonomy and pin yourself down to a particular level. of measurement. The number of categories that comprise the scale may be important. For example, 10-point scales tend to "continuize" things- more than 5-point scales. But even then you have to be careful. (O'Brien [1979] has shown that the dependence of agreement between order-preserving transformations upon the number of categories is not a simple one.) The following considerations might also be useful:

(a) Do you have anything for your raw score scale that even remotely resembles an actual unit of measure (It doesn't have to be in the National Bureau of Standards, but if a subject obtains--a.--score of 3, you should be able to say 3 what.)

(b) Does the scale have a zero point, however arbitrary it may be?

(c) What transformations, if any, of your scale are acceptable? All order- preserving transformations? Only linear transformations? No transformations whatsoever?

4. As far as descriptive statistics is concerned, think about meaningfulness (Suppes, 1959; Marcus-Roberts &Roberts, 1987) before you summarize the data you have in hand, whether for a full population, a random sample, or a convenience sample. If you have to , forgo. traditional statistics such as means, standard deviations, and Pearson r's there are always Tukey's (1977) very creative exploratory data analysis (EDA) techniques, and Agresti (1984) has written a whole book on the analysis of ordinal data. If you are-firmly convinced that you are dealing with ordinal data, you may not be be able to study interactions in the traditionally, much less test their statistical significance. Interactions explicitly or implicitly involve subtraction, which is usually not appropriate for ordinal scales. Multivariate analyses will be at best awkward, and you may not be able to Use your SPSS, SAS, or BMDP computer packages, but computer packages have played the role of master rather than slave far too often.

5. Having made certain decisions regarding scale type and data description, the choice off inferential procedures, if any, should come rather easily. Once again the concept of meaningfulness, or more particularly, the notion of a test of a meaningful hypothesis (Marcus-Roberts & Roberts, 1987), should be kept in mind. If you have decided at the psychometric stage that your scale is ordinal, you are likely to employ some sort of nonparametric test at the inference stage, only because of the distribution-free nature of such tests, but because they tend to be more appropriate for hypotheses that are meaningful for ordinal variables. If you claim that-you have an interval scale, you are more likely to prefer parametric techniques, but should you have qualms about normality and/or homogeneity of variance and elect some nonparametric counterpart, don't be apprehensive about losing power; it maybe even-higher.

One measure of the success of this or any other attempt to resolve the ordinal/interval controversy whether people such as Gaito (1980) and Townsend and Ashby (198.1) can sit down and talk things over, rather than writing disagreeable articles about each other. The "empirical robustness" arguments of Baker et al. (1966), Labovitz (196 7), and their supporters are no longer very convincing, and the most recent arguments do indeed favor Stevens's original position. But a retreat to Siegel's (1956) ultra-conservative stance is not called for. Its time for a truce.

Accepted for publication June 5, 1989.

The author expresses thanks to Fred S. Roberts, PhD. Department of Mathematics. Rutgers University for his very helpful comments regarding previous versions of this paper.

Thomas R. Knapp *EdD is a professor of education and nursing at the University of Rochester, Rochester, .41.*

References

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

Aghesti A. (1984). Analysis of ordinal categorical data. New. York: John Wiley & Sons.

Armstrong, G. D. (1981). Parametric statistics and ordinal data: A pervasive misconception. Nursing *Research, 30,* 60-62.

Armstrong, G. D. (1984). Parametric statistics (Letter to the editor) Nursing *Research, 3,3,* 54.

Baker B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens's proscriptions on statistics. *Education.* al and Psychological *Measurement, 26,* 291309.

Bt-ittt, R. C.. & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. Journal of Educational Statistics, 5 309-335.

Blair, R. C., & Higgins J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. Psychological *Bulletin, 97,* 119-128.

Borgatta, E. F., & Bohrnstedt G. W. (1980). Level of measurement: Once over again. Sociological *Methods and* Research, 9, 147-160.

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

Gardner, P. L. (1975). Scales and statistics. Renew of Educational *Research, 45,* 43J7.

Havelicek, L. L., & Peterson, N. L. (1974). Robustness of the t test: A guide for researchers on effect of violations of assumptions. Psychological *Reports, 34,* 10951114.

Knapp, T. R. (1984). Parametric statistics (Letter to the editor]. Nursing *Research, 33,* 54.

Labovitz, S. (1967). Some observations on measurement and statistics. Social Forces, 46, 151-160.

Marcus-Roberts, H. M., & Roberts, F. S. (1987). Meaningless statistics. Journal of *Educational Statistics*. 12, .38.3-394.

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

Michell, J. (1986). \Ieasurement scales and statistics: A clash of paradigms. Psychological Bulletin 100. 398--107.

O'Brian, R. M. (1979). The use of Pearson's r with ordinal data. American Sociological Review, -1-1. 851-857.

Siegel, S. (1956). Nonparametric statistics *for* the *behavioral sciences.* New York: McGraw-Hill Book Co.

Stevens, S. S. (19-16). On the theory of scales of measurement. Science. 103, 67 t-680.

Stevens, S. S. (1955). On the averaging of data. Science. 121. 11:3-116.

St ppes, P. (1959). Measurement, empirical meaningfulness. and three-valued logic. In C. \t'. Churchman & P. Rat005h (Eds.). Measurement: Definitions and theories (pp 1'2'9-14:1). New York: John Wiley & Sons.

Thomas, H. (1982). IQ interval scales, and normal distributions. Psychological *Bulletin.* 91, 198-202.

Townsend, J. C., & A11111), F. G. (198-t). Measurement scales and statistics: The misconception misconceived. Psychological Bulletin 96, 39 t--101.

Tit achtman J. J., Giambalvo V., & Dippner R. S. (19,8). On the assumptions concerning the assumptions of a t test. Journal of General Psychology, .99. 107-116.

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.