SOME ISSUES IN THE MEASUREMENT-STATISTICS CONTROVERSY

JOHN GAITO
York University

ABSTRACT

Some problems generated by Stevens's pronouncement that measurement scales (nominal, ordinal, interval, ratio) determine specific statistical procedures are discussed. It appears that proponents of this view may think of the statistical analysis stage in research design as equivalent to the overall design process, or that the interpretation stage is included in the statistical analysis stage. This aspect leads to the introduction of irrelevant empirical considerations within conclusions emanating from a statistical analysis. Such pronouncements are faced with certain logical inconsistencies. For example, two or three procedures having different scales yield similar results. Within a statistical analysis there are different contexts or levels of number analysis of different scale nature; yet these differences are not considered in the Stevens approach. Furthermore, the Stevens admonitions can impede progress with theoretical and/or empirical problems. An example is provided in the intelligence measurement area to indicate how important developments have occurred when these admonitions were ignored.

The controversy as to the independence or the non-independence of measurement properties in a statistical analysis is one that generates much intellectual stimulation, but wastes many journal pages. It began in the late 1940s and continued to the early '60s. For over a decade there were an abundance of papers or books either accepting or rejecting the thesis of Stevens (1946) that the specific measurement scale involved with data (nominal, ordinal, interval, ratio) determines the specific operations of a statistical analysis. For example, Burke (1953), Lord (1953), Kaiser (1960), and Anderson (1961) were some of those who claimed independence of measurement and statistical analysis operations. In 1980 Gaito summarized the position of those who opposed the Stevens thesis (hereafter referred to as the AM group — anti-measurement). On the other hand, the major proponents of Stevens's argument (in extreme form) were Siegel (1956) and Senders (1958). Since that period, the issue has been less prominent, but it surfaces periodically, especially in recent years in elementary psychological statistics books (e.g., Horvath, 1985; Pagano, 1981; Walker, 1985). Recently the pro-Stevens approach (hereafter referred to as the PM group — pro-measurement) has been championed by Townsend and Ashby (1984). These authors presented the usual arguments offered by PM personnel, that is, that statistical analyses involve measurement aspects. For example, for an ANOVA situation, PM individuals would require that the data conform to an interval scale, in addition to the usual three assumptions relative to random errors (normal distribution, independence, homogeneity of variance) as expressed in the statement "The e's are NID (0, σ_e^2)." Furthermore, Townsend and Ashby maintain that for the AM group, "essentially 'anything goes' relative to measurement stipulations" (p. 394).

These statements are complemented by comments that I have heard within various settings to the effect that the AM position takes meaning out of research aspects, implying that the AM group are not interested in the relationship between numbers and the underlying referents.

The purpose of this paper is to support the AM approach and to show that "anything does not go" — that there are specific requirements for measurement aspects within a research effort. That is, although measurement considerations do not determine the choice of statistical tests, there are other ways in which measurement aspects are important in the overall experimental design.

Furthermore, the statement accusing the AM group of nonconcern with meaning shows a lack

The author thanks Arnold Binder for reading an early draft of this paper and providing a number of useful suggestions.

Requests for reprints should be addressed to Dr. John Gaito, Department of Psychology, York University, Downsview, Ont. M3J 1P3.

of understanding of what those in this group are stating. I suggest that one main difference between the AM and PM positions is that the term "statistical analysis" has a different meaning and emphasis for the two groups. *The AM group is defining or emphasizing statistical analysis as one of a number of stages within experimental design, but the PM group may be equating statistical analysis with the overall research effort, or confounding the interpretation phase with the statistical analysis operations.* Implicit within this aspect is a second point of difference, that of statistical conclusions and empirical conclusions. Each of these will be considered. Then some serious logical inconsistencies and empirical shortcomings of the PM position will be discussed.[1]

Experimental Design vs. Statistical Analysis

Experimental design can be conceived of as involving four stages in an overall research effort. More stages might be suggested, but four will suffice for the purpose of this article. These four stages are: planning and design of the experiment; conduct of the experiment; statistical analysis of data; interpretation of results.

It is possible that the PM group is concerned with the overall experimental design stages (or the last stage — interpretation) when they talk of measurement properties as being important considerations in statistical analyses (the third stage). However, there are important distinctions between the operations involved in each of the separate stages, relevant to measurement aspects.

The AM group would agree that measurement properties are important in the overall experimental design, specifically in the planning/design stage and in the interpretation of results. No AM member would dispute that fact that measurement considerations such as reliability, validity, relevancy, and (especially) meaningfulness of the dependent variables should enter into consideration during the planning stage. Furthermore, these measurement aspects are important in making sense (i.e., providing meaning) of the results in the interpretation stage. It would be unrealistic not to accept these measurement notions, for the research effort would be meaningless.

However, the AM group would not equate statistical analysis with experimental design; their ideas refer only to the statistical analysis stage. This stage is merely one in the overall research effort in which mathematical operations hold sway and measurement scale considerations are irrelevant. This is the domain to which the many papers of the AM group are directed. (See, for example, the excellent papers by Burke, 1953, and Lord, 1953, which should have settled the problem over three decades ago.)

Let us look closely at a statistical analysis as viewed by the AM group. This stage is concerned with analyses involving events such as determining medians or means, variances, correlations, etc. and with tests of null hypotheses (H_0). In the latter case, the observed results are contrasted with the values expected based on specific mathematical assumptions that are present in the mathematical model for the procedure. The investigator then decides whether to reject, or not reject, H_0 on the basis of a specific probability level. With the decision to reject, or not reject, H_0, the statistical conclusion is that there is one, or more than one, population distribution from which the samples have been chosen. These are statistical conclusions that emanate from the mathematical operations involved in the specific procedure. These conclusions are completely devoid of empirical aspects (i.e., those inherent in the experimental and theoretical nature of the research effort), and characteristics of data such as reliability, validity, meaningfulness, and relevancy do not enter the picture. Specifically, as Lord implied, "The numbers do not know where they came from."

Likewise, the conduct of the experiment is a physical act and measurement aspects are irrelevant. But this stage is of little consequence for this paper.

Statistical Conclusions vs. Empirical Conclusions

Another point of contention that seems to be present in the controversy is the possible confusion between conclusions of a statistical nature and those falling in the empirical domain. As indicated in the last section, statistical conclusions are defined within the mathematical context of the procedure and follow the decision to

[1] A detailed discussion of formal measurement theory is not the intention of this paper. An excellent detailed discussion is provided by Binder (1984). For more comprehensive formal treatments, see Adams, Fagot, and Robinson (1965) and Pfanzagl (1968).

reject H_0 or not. For example, let us take an ANOVA situation; the conclusion in the case wherein rejection of H_0 occurs is that the samples come from two or more population distributions. The conclusion is that at least one set of numbers is different from other sets. There is no concern with what the numbers refer to. The set of numbers is merely a distribution of values. This is a *gross statistical statement*, that the two or more samples are from different population distributions. In the case in which H_0 is not rejected, the conclusion is that there is no evidence to indicate that the samples come from more than one population distribution. This also is a *gross statistical statement*, that there is no evidence to indicate that the populations from which the samples were derived are different; that is, only one population distribution is involved. In statistical analyses there is no reference to measurement aspects involved in the numbers. *The conclusion is that the populations are different, or not.*

On the other hand, empirical or theoretical conclusions do have reference to what the numbers stand for and mean. Thus measurement properties enter the picture. The researcher takes the results of the statistical analysis stage and places them within the context of the research effort. For example, if one is conducting a learning experiment and is using number of errors as an indicator of degree of learning, measurement considerations arise concerning this choice, for example, reliability, validity, meaningfulness, and relevancy of the index. If these aspects were handled in an adequate fashion before the experiment was conducted, then the researcher can conclude that different degrees of learning occurred in the specific situation of concern (if H_0 was rejected) or that there is no evidence to indicate different degrees of learning (if H_0 was not rejected). The interpretation stage brings in the specific empirical operations with their associated measurement requirements so as to provide meaningful interpretation of the results of the experiment.

In summary, 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 the interpretation stage.

Logical Inconsistencies in the PM Position

The position of the PM group — that measurement scale properties of the data determine the specific statistical analyses — encounters a number of logical inconsistencies. These are:

Levels of Number Analysis or Context

The context of the number analysis in which the assignment of the measurement scale property occurs is of major importance (Gaito, 1960). There can be more than one specific level of number analysis involved in this assignment. For example, take the case wherein a number is given to a single response of one subject on one occasion: *S* gives a response to a test item and is scored right (1) or wrong (0). The number of 1 or 0 in this case would indicate the lowest scale level, nominal data, according to the PM group.

However, if we determine the total number of correct responses of one subject, then a different scale should appear. This scale is at least an ordinal one; for example, 20 correct of 20 responses is greater than 19 correct in 20 items. The same result would occur if there were more than one subject. Likewise, if the mean or median of the set of scores for one subject (or more than one subject) were determined, at least an ordinal scale would appear.

Finally, if these correct responses are considered as a sample drawn from a population distribution of correct responses and the characteristics of this distribution are determinable, then an interval scale is involved — the differences between various points on the curve can be of known value.

In this example, we have demonstrated three different contexts or levels of number analyses that can be present in a statistical analysis. It appears that the PM group has been concerned only with one of these levels in each case. Yet in a statistical analysis all three contexts can appear. Even with the use of a χ^2 test of the null hypothesis, which the PM group would specify as an example of nominal scale statistics, all three levels are involved. One could ask why there is concern with only the first level in this case when this statistical procedure involves also the obtaining of the frequency of responses (level 2 — a frequency of 10 is greater than a frequency of 9, etc.) and relating these obtained frequencies to the expected frequencies using a familiar distribution (χ^2 — level 3 analysis).

Different Scales Give Same Result

A serious problem with the notion that the measurement scale of data determines the statistical procedures has been pointed out by a number of writers (e.g., Binder, 1984; Gaito, 1980; Savage, 1957). This is the logical inconsistency involved when two or more procedures exemplify different types of scale properties but produce the same result. Three examples should be sufficient.

(a) The Binomial Test and the Sign Test are supposed to consist of nominal data and ordinal data, respectively. However, underlying both techniques is the binomial distribution and both allow for the rejection, or non-rejection, of H_0 at the same probability level.

(b) *The normal distribution* (interval scale) provides an excellent approximation to the exact probabilities given by the *binomial distribution* (nominal scale, according to PM group), especially when $p = q$ and n is 10 or more.

(c) With classificatory data, χ^2 (nominal scale) and the *normal distribution* (interval scale) can give the same result under some conditions. This is to be expected since the square of a unit normal variate has a chi-square distribution with 1 *df*, i.e., $z^2 = \chi^2$ when 1 *df* occurs.

Actually (b) and (c) examples can be combined. *In some cases the results of the use of* χ^2*, the binomial distribution, and the normal approximation provide similar results.*

The first example illustrates data that have adjacent scale properties (nominal — ordinal). However, the second and third ones would appear to be the most difficult ones for PM advocates to handle, because these examples involve data that are two scales apart (nominal — interval). Unfortunately, members of the PM group have not noted or commented on this apparent inconsistency. In any event, these three examples should indicate clearly the independence of measurement scales and statistical analyses. In actual fact, there are only two types of data, continuous and discontinuous. However, in some cases even this distinction becomes blurred — for example, when the normal distribution *(continuous in form) is used as an approximation* to the discontinuous distributions cited above.[2]

Other Examples

Another example of logical inconsistency in the PM position is cited by Binder (1984). He describes an example in a book by Johnson (1981). The latter specifies that use of Pearson's *r* requires interval (or ratio) type data. Johnson then adds that the *rho is the Pearson correlation for the same data in ranks*, and is of ordinal scale nature, and he apparently did not recognize the inconsistency involved. However, Spearman's method is a type of product moment procedure that provides simplified calculations that depend on the numerical properties of ranks; that is, Spearman's rho is an estimate of the Pearson *r* when the numerical values of the latter are converted to ranks.

Furthermore, the PM group allows for only certain transformations (permissible) to occur with each measurement scale. Yet a number of researchers have shown clearly that non-permissible transformations of data produce similar results to those with permissible transformations, indicating that statistical tests depend upon numbers and not their histories or source (Anderson, 1961; Baker, Hardyck, & Petrinovich, 1966; Binder, 1984).

Also, in many cases there is the statement or implication that the operations of addition, subtraction, multiplication, and division cannot occur with subinterval type data (and that nonparametric procedures should be used). To which one can question, as for example did Lubin (1962, p. 359),

> How does one compute chi-square, Spearman's rho, Wilcoxon's U, or any other nonparametric statistic without adding, subtracting, multiplying, or dividing?

In summary, it seems that the PM group either have not recognized the many inconsistencies in their position or else have superficially cast them aside. The latter aspect occurred on a number of occasions in the paper by Townsend and Ashby (1984). For example, in response to the Gaito article (1980) they state "we were simply not

[2] See Binder (1984) and Savage (1957) for additional examples of inconsistencies.

able to make sense of. . ." (p. 395). In commenting on the central point implied by Lord (1953) in his much cited paper "that the numbers do not know where they came from," they indicate that "just exactly what this curious statement has to do with statistics or measurement eludes us" (p. 396). The statements of Savage (1957) "seem beside the point" (p. 396). It appears that if they do not understand a point, it must be incorrect. Yet another interpretation of these responses could be that the authors are showing a lack of understanding of the basic points involved.

Theoretical and Empirical Shortcomings in the PM Position

It should be emphasized that measurement and statistical procedures are tools that the scientist uses to attain certain empirical and theoretical objectives. Thus, the scientist should make use of any tools that will facilitate movement toward the goals. The consequences of following slavishly the pronouncements of the PM group can result in the loss of potential theoretical and empirical gains. For example, Binder (1984) indicated that by disregarding the suggestion that IQ is measurable only on an ordinal scale,

> . . . investigators computed means and standard deviations with I.Q.'s, correlated I.Q.'s with many other variables (some of which were nominal), and tested hypotheses involving the I.Q. with analysis of variance. What resulted was rich, empirical knowledge, a theoretical structure that matches any other structure in the social sciences for predictive usefulness . . . . The point is that important empirical advances were made by procedures that were said to be inappropriate by Stevens, Siegel, and the others. (p. 475).

This example shows not only that the admonitions of the PM group can impede theoretical-empirical developments in an area of science, but also that the relating of statistical analyses results to the empirical domain often precedes, and may indeed lead to, ultimate determination of measurement properties (Binder, personal communication).

Conclusions

A central point in the argument of the PM group is that the AM group does not allow measurement aspects with associated meaning to enter into the overall research picture. As the above discussion indicates, *this is not a correct description of the AM approach*. The AM group allows measurement and meaning to enter into some stages of experimental design (i.e., planning, interpretation of results), but not into others, specifically not into statistical analyses. Only mathematical, not measurement, aspects enter at this point. This is the point on which the excellent early papers by Burke and by Lord are based.

Furthermore, the logical inconsistency and empirical shortcomings of the examples involved in the last two sections would appear to be difficult to rationalize by the PM group. However, it seems that members of this group overlook these types of possibilities. Binder (1984) indicated that such inconsistencies occur because PM advocates miss the point that "levels of measurement" refers to relationships between empirical and numerical worlds, rather than being intrinsic characteristics of numbers.

It is interesting, but not unexpected, that not only psychologists are confused by this issue; the problem permeates other disciplines as well, such as international relations (R.J. Rummel, personal communication[3]) and criminology (Binder, 1984). However, as indicated in the above discussion, the argument between the AM and PM groups might be alleviated if both groups were to recognize statistical analysis, with its specific operations, as merely one of a number of stages in the overall research effort. It is obvious that both measurement and statistical considerations are involved in all specific experimental designs.

RÉSUMÉ

Ce travail présente certains des problèmes soulevés par les déclarations de Stevens assurant que les échelles de mesure (nominales, ordinales, à intervalles, à proportions) déterminent des procédures statistiques particulières. Ceux qui partagent ce point de vue semblent penser que l'étape de l'analyse statistique dans le plan de recherche équivaut au processus du plan d'ensemble, ou semblent inclure le stade de l'interprétation à celui de l'analyse statistique. Ils sont ainsi amenés à inclure dans des conclusions qui proviennent d'analyses

[3] R.J. Rummel, Department of International Relations, University of Hawaii, Honolulu, Hawaii.

statistiques des considérations empiriques non-pertinentes, ce qui produit certains illogismes. Par exemple, deux ou trois procédures qui ont des échelles différentes donnent des résultats semblables. Les analyses statistiques comportent des contextes différents ou des niveaux d'analyse numérique dont les types d'échelles sont différents. Or l'approche de Stevens ne prend pas ces différences en considération. Par ailleurs les admonestations de Stevens risquent de freiner la résolution des problèmes empiriques et/ou théoriques. L'auteur présente un exemple montrant qu'il s'est produit des développements importants lorsqu'on avait ignoré ces admonestations : dans le domaine des mesures de l'intelligence.

References

Adam, E.W., Fagot, R.F., & Robinson, R.E. (1956). A theory of appropriate statistics. *Psychometrika, 30,* 99-127.

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's proscriptions on statistics. *Educational and Psychological Measurement, 26,* 291-309.

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.

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

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.

Horvath, T. (1985). *Basic statistics for behavioral sciences.* Boston: Little, Brown and Co.

Johnson, E.S. (1981). Research methods in criminology and criminal justice. *British Journal of Criminology, 10,* 124-135.

Kaiser, H. (1960). Review of Senders' *Measurement and statistics. Psychometrika, 25,* 411-413.

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

Lubin, A. (1962). Statistics. *Annual Review of Psychology, 13,* 345-370.

Pagano, R.R. (1981). *Understanding statistics in the behavioral sciences.* New York: West Publishing Co.

Pfanzagl, J. (1968). *Theory of measurement.* Würzburg-Wien: Physica Verlag.

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

Senders, V.L. (1958). *Measurement and statistics.* New York: Oxford University Press.

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

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.

Walker, J.T. (1985). *Using statistics for psychological research: An introduction.* New York: Holt, Rinehart and Co.