Bulletin of the Psychonomic Society
1978, Vol. 11 (2), 136-138

The utility of utility indices

DAVID R. SODERQUIST and RICHARD A. HUSSIAN
University of North Carolina, Greensboro, North Carolina 27412

Many researchers have questioned the incorporation of utility indices following significant results yielded by the F ratio. Problems frequently mentioned concerning magnitude estimates include biases, questionable model utilization, acceptable size of the UI, and effects of sample size. These issues are addressed and arguments for the inclusion of a magnitude estimate in report writing are tendered. Results of a frequency study on the use of utility measures are included, which show that very few published reports incorporate magnitude estimates.

Once there was a psychologist who was interested in investigating variables that affect the behavior of sailors, fishermen, and retired psychologists. The psychologist, Dr. Effect, was interested in predicting the tidal cycle of the coastal waterways, although he had absolutely no knowledge concerning the movements of celestial bodies and tides. However, this did not impede his activity into the unknown. He simply began his investigation by selecting particular planets and looking for their effect on the tide. Since Dr. Effect had, as a graduate student, been subjected to a rigorous training in statistics and quantitative methods, he naturally decided on a statistical approach to the problem. That is, he analyzed reams of data (collected, of course, by a multitude of graduate students). The final analysis yielded the conclusion that Venus, Mars, and Jupiter all had an effect on the tides. Moreover, the probability that the observed results were due to chance was less than 1 in 10,000. The results were so overwhelming that Dr. Effect rushed the finding into print. His reward was instant tenure and an enormous grant from NSF to continue his highly significant research. Although Dr. Effect accepted the glory and recognition, he realized that he really had not helped the sailors, fishermen, and retired psychologists very much. Going to the beach to launch their boats, particularly for the retired psychologists, was still largely a random event. They simply could not apply the highly significant results to reality. The data had little practical use in terms of predicting the status of the tides. Then one day, at coffee, a graduate student happened to suggest to Dr. Effect that he compute a statistic or utility index (UI) that would estimate the strength of the relationship between the celestial bodies and the tide. Such an index would quantify the usefulness of a significant effect. According to the graduate student, who was studying the well known text by Hays (1963), all that was necessary was to compute the statistic omega square (or eta square, or epsilon square, or possibly the Pearson PM coefficient square). At the graduate student's urging, Dr. Effect reviewed the literature and found that strength of effect indices had been in existence for over 60 years. After recovering from his initial surprise, Dr. Effect calculated the omega square index suggested by Hays and found that the significant result he had so proudly pronounced to the world accounted for less than .001% of the total variation in the experiment. Obviously, the data were significant but, based on omega square, they had no utility in terms of prediciton. It was then clear why the sailors, fishermen, and retired psychologists could not use the data to predict the tidal rhythm. Omega square clearly revealed the futility of using the data to make any predictions. The relationship between the celestial bodies and the tide was highly significant, yet completely worthless in terms of utility and prediction.

This fable illustrates a serious problem in the psychological literature; namely, significant effects with little predictive utility. It is exactly this problem which Hays addressed in his text when he suggested the use of omega square. On the surface, the problem appears to have a simple solution. namely, determine statistical significance with the usual F ratio and then estimate the magnitude of the relationship with a statistic such as omega square.

This solution, although simple, apparently has not made a lasting impression on psychological investigators.[1] Moreover, some researchers have suggested that such indices are unwarranted or misleading (Dooling & Danks, 1975; Glass & Hakstian, 1969). It is our intent, therefore, to briefly examine some of the common problems thought to be associated with an index of utility.

Since there are several means of calculating the strength of a relationship, we have restricted ourselves to an estimate that is based on the expected mean squares (EMS) associated with the ANOVA model. We refer to this quantity as a utility index (UI), following the suggestion of Bolles and Messick (1958). The derivation and computational formulae for this index have been explicitly spelled out by Dodd and Schultz (1973), Gaebelein and Soderquist (in press), Gaebelein, Soderquist, and Powers (1976), Gaito (1958, 1960), and Vaughan and Corballis (1969). Although there are a few questions that relate to these derivations, they do not generally cause concern among the potential users of such an index. Our focus, then, is not on derivations;

rather, it is on the criticisms and confusion that surround the interpretation of UIs. An examination of the literature indicates that there is little interpretation problem when the underlying ANOVA model is assumed to be random. That is, in the random model, there is external validity for significant effects and the UI. Just as one can generalize to levels of the independent variable other than those employed in the study given a significant F ratio for a random variable, so can the magnitude of the UI for the effect be generalized. This allows one to quantify the importance of the independent variable (IV) between and within studies.[2] The literature, however, clearly shows that the interpretation of UIs is uncertain when fixed effects are assumed. Since the majority of the investigations done by psychologists assumes a fixed effects model, our comments are specifically directed at this model.

RANGE OF SELECTED VARIABLES

There are two points to be considered here. First, other things being equal, a UI and an F ratio varies with the investigator's choice of levels of the IV. Extremely divergent levels of the IV are more likely to yield a significant F and a large UI than are levels of an IV that are in close proximity to each other. Second, if an investigator selects all possible levels of the IV, the external validity becomes irrelevant, since the entire population is represented in the study. In both of these situations, the UI is a descriptive index representing the percentage of variability in the experiment that is associated with the selected levels of the IV. There is no external validity for the UI, nor for the significant effects obtained in the experiment. Contrasting UIs across separate investigations is not valid for the same logical reason that the comparisons of F ratios are not valid. There is no external validity unless the experiments are identical. Furthermore, the contrast of UIs within an experiment is restricted and can be interpreted only in terms of the specific variables employed.

MONTE CARLO STUDIES

Hays (1963) proposed the measure of utility (ω^2) for two reasons. First, with small sample sizes a reasonably large UI would supposedly indicate a useful relationship, in spite of a nonsignificant F ratio. Second, a UI could be applied to data where a significant F ratio may have been obtained as a result of a large sample size; hence, the predictive utility of the study could be quantitatively evaluated. Carroll and Nordholm (1975) and Keselman (1975) have recently shown that the usefulness of a UI is limited to situations where sample size is relatively large. The UI estimate, like other statistics, becomes a good estimate as a direct function of the sample size. Therefore, the UI is more useful when N is large. The standard error of the UI is the limiting factor in evaluating the utility of significant and nonsignificant F ratios. When sample size is small, the UI estimate lacks the desired precision. However, the estimate "does have strong merit" with large sample sizes (Carroll & Nordholm, 1975).

MAGNITUDE OF THE UI

The size of an acceptable utility index is another issue related to interpretation. That is, how large must a UI be in order to reflect some practical use? The answer to this question is not straightforward. Investigators have standards against which to judge significance, but there is no standard available for the UI. This is true for at least two reasons. First, the magnitude of the UI will vary as a function of the range of levels within each IV. Since the UI is a proportion of the total variance within the experiment that is accounted for by the particular levels of the IV that have been chosen, it is clear that no criterion analogous to an alpha level is applicable. Second, since psychology is far from making precise predictions concerning behavior, it is likely that the UI will vary across the different subspecialties of psychology. For example, a UI of .15 may be unimpressive in visual psychophysics, while being substantial in a field such as social psychology.[3] In short, the magnitude of the UI will vary as a function of the choice of variables and the particular area of psychology under investigation; hence, the criteria that determines usefulness of the UI is relegated to the expertise of the investigator. This conclusion, however, does not detract from the usefulness of the UI when one considers the fact that, in fixed effects designs, the expertise of the investigator is nearly always used when the levels of the IV are chosen anyway. The UI, in this light, may be a very important quantity in the overall evaluation of the data.

MIXED MODELS

Factorial experiments that have random and fixed variables within the same design require special consideration. The assumptions about the linear model can lead to different estimates of the magnitude of effect. That is, there is no clear consensus among statisticians as to how an interaction effect should be treated; for example, is an interaction assumed to be fixed or random when one variable is fixed and one is random? The reader is directed to Dwyer (1974) and to Gaebelein, Soderquist, and Powers (1976) for further discussion of this issue.

In summary, it is apparent that each investigator must make his own decision concerning data interpretation and analysis. The incorporation of the UI provides additional information and is a useful index when properly interpreted. When an investigator is aware of

the aforementioned cautions regarding the choice of variables, the underlying model, and the inherent biases, the UI can add substantially to the interpretation of experimental results.

REFERENCES

BOLLES, R., & MESSICK, S. Statistical untility in experimental inference. *Psychological Reports*, 1958, **4**, 223-227.

CARROLL, R. M., & NORDHOLM, L. A. Sampling characteristics of Kelley's ε^2 and Hays' ε^2. *Educational and Psychological Measurement*, 1975, **35**, 541-554.

CRAIG, J. R., EISON, C. L., & METZE, L. P. Significance tests and their interpretation: An example utilizing published research and ω^2. *Bulletin of the Psychonomic Society*, 1976, **7**, 280-282.

DODD, D. H., & SCHULTZ, R. F., JR. Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin*, 1973, **79**, 391-395.

DOOLING, D. J., & DANKS, J. H. Going beyond tests of significance: Is Psychology ready? *Bulletin of the Psychonomic Society*, 1975, **5**, 15-17.

DWYER, J. H. Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin*, 1974, **81**, 731-737.

GAEBELEIN, J. W., & SODERQUIST, D. R. The utility of within-subjects variables: Estimates of strength. *Educational and Psychological Measurement*, in press.

GAEBELEIN, J. W., SODERQUIST, D. R., & POWERS, W. A. A note on variance explained in the mixed analysis of variance model. *Psychological Bulletin*, 1976, **83**, 1110-1112.

GAITO, J. The Bolles-Messick coefficient of utility. *Psychological Reports*, 1958, **4**, 595-598.

GAITO, J. Expected mean squares in analysis of variance techniques. *Psychological Reports*, 1960, **7**, 3-10.

GLASS, G. W., & HAKSTIAN, A. R. Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, 1969, **6**, 403-414.

HAYS, W. L. *Statistics*. New York: Holt, Rinehart, & Winston, 1963.

KESELMAN, H. J. A Monte Carlo investigation of three estimates of treatment magnitude: Epsilon squared, eta squared, and omega squared. *Canadian Psychological Review*, 1975, **16**, 44-48.

VAUGHAN, G. M., & CORBALLIS, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, 1969, **72**, 204-213.

NOTES

1. The frequency of magnitude estimate utilization was determined by a survey of six selected journals from 1970 through 1976, including *Journal of Abnormal Psychology, Journal of Comparative and Physiological Psychology, Journal of Experimental Psychology, Journal of Personality and Social Psychology, Learning and Motivation,* and *Memory & Cognition* (1973-1976). Magnitude estimates following t tests and F tests were conducted in approximately 2% of these experiments; no increase in the use of magnitude estimates during the 6-year period was discernible. Craig, Eison, and Metze (1976) also reviewed the literature and found essentially the same result. Recently, however, several journals (e.g., *Cognitive Therapy and Research* and *Learning and Motivation*) have advocated the use of magnitude estimates in their editorial policy.

2. Although Dooling and Danks (1975) have given an example showing how UIs may be misinterpreted if they are contrasted within a particular study, it is equally clear that F ratios may be misinterpreted for the same reason; that is, skewed population distributions (regardless of the ANOVA model). These violations, which may cause unrepresentative UIs, may also yield unrepresentative F ratios, since both of these quantities are dependent upon the same assumptions and EMS. The robustness of the ANOVA suggests that this criticism may not be as relevant as Dooling and Danks imply.

3. This rationale applies to the random model as well as the fixed model.

(Received for publication October 10, 1977.)