Kendler, H. H., Kendler, T. S., & Wells, D. Reversal and nonreversal shifts in nursery school children. Journal of Comparative & Physiological Psychology, 1960, 53, 83-87.
Kendler, T. S. Verbalization and optional reversal shifts among kindergarten children. Journal of Verbal Learning & Verbal Behavior, 1964, 3, 428-433.
McAdam, D. W., & Whitaker, H. A. Language production: Electroencephalographic localization in the normal human brain. Science, 1971, 172, 499-502.
Penfield, W., & Roberts, L. *Speech and brain mechanisms.* Princeton: Princeton University Press, 1959.

(Received for publication August 21, 1974.)

Bulletin of the Psychonomic Society
1975, Vol. 5 (1), 15-17

Going beyond tests of significance: Is psychology ready?*

D. JAMES DOOLING and JOSEPH H. DANKS
Kent State University, Kent, Ohio 44242

A criticism is offered for the recent efforts to encourage psychologists to report the proportion of variance accounted for (ω^2) by analysis of variance effects. Few psychological experiments employ designs that allow legitimate inferences as to the strength of particular effects. As such, ω^2 is a descriptive statistic that is extremely limited in its usefulness. It is suggested that a widespread reporting of ω^2 in psychology is not only unnecessary, but could also be misleading.

In the past decade, some psychologists have criticized the mere reporting of significance levels in analysis-of-variance designs. According to this view, it is also important, even essential, to report the "strength" of the significant effects that have been obtained. In addition to the reporting of F-ratios, psychologists have been urged to compute a statistic that reflects the proportion of variance accounted for by a given effect (ω^2). A source most influential in this regard was Hays's *Statistics for psychologists* (1963). His argument can be summarized as follows: "The occurrence of a significant result says nothing at all about the strength of the association between treatment and score. A significant result leads to the inference that some association exists, but in no sense does this mean that an important degree of association necessarily exists. Conversely, evidence of a strong statistical association can occur in data even when the results are not significant [1963, p. 342]." More recently, Vaughan and Corballis (1969) called attention to the fact that research psychologists have not heeded Hays's advice. They contended that lack of such information in published articles is a serious problem and that "a change of attitude is required [1969, p. 204]." Following Vaughan and Corballis four years later, Dodd and Schultz (1973) echo their cry, providing computational procedures in an attempt to "encourage researchers to include estimates of magnitude of effects [p. 395]."

We take a position in opposition to the opinions expressed above, though we find merit in the positions stated. Our comment does not fault either the reasoning of Hays or the computational formulas provided by either Vaughan and Corballis or by Dodd and Schultz. We find the logic of the Hays comment impeccable and agree that knowledge of the strength of association would add new information not given by the usual report of significance levels alone. Our quarrel is rather with a possible (and we think, probable) misinterpretation of the proportion of variance accounted for measure (ω^2) when reported in journals and read by psychologists. In addition, we argue that the

*We are grateful to Clyde Hendrick, Roy Lachman, Roy S. Lilly, and Terry J. Spencer for their comments on an earlier version of this manuscript.

present state of the art in psychological experimentation vitiates calculation of such a statistic in most instances.

The crux of our argument is that in the experimental designs typically employed by psychologists, the selection of the levels of the independent variable is not sufficiently well defined to permit legitimate inferences about the strength of that independent variable's effect. In most psychological experiments, the experimenter is interested in whether or not a particular independent variable has any effect at all on his dependent variable. He, therefore, selects two or more levels of the independent variable for experimental manipulation. In this selection process, the factors controlling the selection are not always obvious. If the experiment is an initial one on a problem, the E will likely pick extreme values in order to maximize the probability of obtaining a significant effect. Or the E may be restricted to the materials used in some prior experiment that he is replicating or extending. Or he may be constrained by the values easily manipulated on his apparatus. In each of these cases, the E typically employs the fixed-effects model in his analysis of variance because he did not randomly select the levels of the independent variable.[1] In attempting to assess the experimental effect, the E cannot then legitimately generalize the results of his analysis to all conceivable levels of the independent variable. Consequently, ω^2 appropriately describes only the strength of the *particular* levels of the independent variable on the dependent variable, not the influence of the independent variable *in general.*

How would a researcher be likely to interpret an ω^2 that is obtained from a typical experiment? There would be no problem if ω^2 were read as a purely descriptive statistic, limited to the characteristics of one particular set of data. But he will undoubtedly generalize the statistic beyond the specific levels employed by comparing the ω^2 with those of other independent variables either within the same experiment or from other experiments. For such purposes of statistical inference, ω^2 is either useless, misleading, or both because the strength of an experimental effect will depend directly on ill-defined decisions by the E. The E who chooses extreme values of his independent variable is likely to obtain strong effects (assuming, of course, that there is some relationship between independent and dependent variables). Another E might choose closely related values of the same independent variable that would in turn yield a smaller ω^2. Hence, a correct inference based on the ω^2 is impossible because the underlying population of values on the independent variable has not been randomly sampled. Since most readers of research articles are likely to draw unwarranted inferences from a reported ω^2, we suggest that its widespread use be discouraged, contrary to the conclusion of Vaughan and Corballis (1969) and Dodd and Schultz (1973).

It is clear that the problems addressed in this paper are not formally statistical. They have to do rather with the limitations of the experimental designs commonly employed in psychology. Why not, then, keep the statistic, ω^2, and change the design practices of psychologists? There is, of course, room for improvement. For example, the use of the random effects model should be encouraged when the researcher is dealing with an independent variable whose population characteristics are fairly well understood. But in most cases, major change is not possible. Psychology is a science that still does not know what its major independent variables are. Hence, most experiments tend to be limited to hypothesis testing, as opposed to parameter estimation. Few variables are reliable enough in their effects to permit extensive parametric research. In order to make the ω^2 statistic useful for comparisons both within and between experiments, much standardization would have to occur. But most psychologists would rightly resist any effort in that direction. We are simply not ready to force ill-defined variables into standard molds. Nor would there be any wide agreement on standard experimental procedures in the various areas of psychology. Such efforts at standardization would eliminate many potentially interesting variables that would now be considered "extraneous' and thereby limit the scope of psychological research. Clearly, there is no realistic hope that experiments can be redesigned in order to make ω^2 a useful measure.

The points made thus far may seem to be elementary. Indeed, the difference between a descriptive and inferential statistic is covered in every undergraduate statistics course. Nevertheless, we feel that the potential for misinterpretation of ω^2 is very great. Two personal examples may help to illustrate this point.

One of us (DJD) is currently engaged in some research, as yet unpublished, on the comprehension speed of nominalization phrases like *growling lions* and *raising flowers*. In the course of investigating the effects of two independent variables (grammatical complexity and rated imagery) on comprehension speed, it was noticed that a third variable, transitional probability, might be having an important influence on the comprehension times. A post hoc correlation was computed between transitional probability and comprehension time for all phrases used in the experiment. The result was a statistically significant ($p < .05$) correlation of $-.26$. But because transitional probability accounted for less than 7% of the variance, it was assumed to be only weakly related to comprehension speed and not a variable of major interest. Nevertheless, in spite of the statistical evidence, the variable was followed up in a new experiment. Two lists of nominalization phrases were created that were equal in grammatical complexity and ratings of imagery, but which differed extremely in transitional probability. In a fixed effects analysis of variance, the main effect of

transitional probability was highly significant at $p < .0001$. More relevant to the present purposes, this independent variable now accounted for a much larger share of the variance, $\omega^2 = .23$. A variable that accounts for 23% of the variance tends to be taken more seriously by a researcher than one that accounts for only 7%. Now, which of the two conflicting estimates of the strength of transitional probability is the "correct" one? The answer has to be "neither." For in neither case were procedures followed to insure that levels of transitional probability were randomly sampled, thus, allowing an unbiased statistical inference. The tests of significance leave us confident that transitional probability has an effect on comprehension speed. The design of the experiments, however, prevents us from going beyond this conclusion to make a statement on the strength of the transitional probability variable.

A second example is derived from the work of Danks (1969). He varied the grammaticalness and meaningfulness of sentences and assessed the effect of these variables on comprehension time. Each variable was dichotomous: grammatical vs ungrammatical sentences and meaningful vs nonmeaningful sentences. Both variables had statistically significant effects on comprehension time. In addition, meaningfulness accounted for approximately twice as much variance (ω^2) as did grammaticalness. Danks concluded that semantics was a more important process in the comprehension of sentences than was syntax. While this conclusion may well be true, one cannot draw that conclusion with any assurance from the experiment in question. The manipulation of grammaticalness and meaningfulness may not have been accomplished in a comparable way for both variables. The sentences were constructed by the author in such a way that Ss rated them as differing along the two dimensions of grammar and meaning. But one cannot be sure that the relative differences as manipulated in this experiment matched the relative differences between grammar and meaning in the population of all possible word strings. The fact that ω^2 can be easily misinterpreted is well illustrated by this example. The unwarranted conclusion on the relative importance of syntax and semantics was accepted by a doctoral dissertation committee and by the editor of a widely respected journal.

A recent experiment by Loftus (1973) illustrates an important implication of our argument. In the design of her experiment, she was evidently sensitive to the fact that you cannot arbitrarily choose the levels of two variables and then directly compare the strengths of their effects. The two variables of interest were (a) category dominance; the probability with which a particular category is given as a response to an instance, e.g., *insect* as a response to *butterfly*, and (b) instance dominance; the probability with which a particular instance is given as a response to a category, e.g., *shrimp* as a response to *seafood*. From normative data, Loftus chose instances of high and low dominance for both category and instance dominance. In so doing, she used the same probabilities as cutoffs on each variable to define high or low values. High dominance responses in both instances were those given more than 70% of the time by the normative Ss. Low dominance was reflected in responses that occurred with a frequency less than 26%. Both variables led to significant effects on reaction time, but Loftus also computed ω^2 to decide the relative importance of the two variables. Superficially, this may seem to be a legitimate procedure, but it has a potential flaw. The underlying distributions of the two variables may have important differences. For example, we suspect that the distribution of category dominances is more likely to be negatively skewed than the instance dominance distribution because Ss are more likely to agree about categories than about instances. Because of this potential inequality of distributions, this comparison of ω^2s was inappropriate. In fairness to Loftus, we should add that her conclusions do not rest solely on this particular statistical comparison. Other data from her study provide independent verification for her conclusions.

In conclusion, we doubt the possibility of going "beyond tests of significance" in the analysis of most psychological experiments. The questions to be answered and the statistics to be reported will depend on the type of experimental design employed. By far, the greatest number of research designs used in psychology limit the scope of the experiment to the question: "Does this variable have an effect?" A test of significance is entirely adequate for answering such a question. Most analysis of variance experiments in psychology are not designed to provide information on the further queston: "How strongly are these two variables related?" In such experiments, to report the proportion of variance accounted for will in all likelihood mislead both the Es and the readers of the research.

REFERENCES

Danks, J. H. Grammaticalness and meaningfulness in the comprehension of sentences. Journal of Verbal Learning & Verbal Behavior, 1969, 8, 687-696.

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

Hays, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

Loftus, E. F. Category dominance, instance dominance, and categorization time. Journal of Experimental Psychology, 1973, 97, 70-74.

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

NOTE

1. The random selection of *levels* is different from the random selection of stimulus items within each level. Psychologists typically go to great lengths to accomplish the second randomization; however, such efforts do not solve the problem we are discussing.

(Received for publication September 3, 1974.)