Br. J. math. statist. Psychol. (1978), **31**, 107–112 *Printed in Great Britain*

Estimating effect size: Bias resulting from the significance criterion in editorial decisions

David M. Lane and William P. Dunlap

Experiments that find larger differences between groups than actually exist in the population are more likely to pass stringent tests of significance and be published than experiments that find smaller differences. Published measures of the magnitude of experimental effects will therefore tend to overestimate these effects. This bias was investigated as a function of sample size, actual population difference, and alpha level. The overestimation of experimental effects was found to be quite large with the commonly employed significance levels of 5 per cent and 1 per cent. Further, the recently recommended measure, ω^2, was found to depend much more heavily on the alpha level employed than the true population ω^2 value. Hence, it was concluded that effect size estimation is impractical unless scientific journals drop the consideration of statistical significance as one of the criteria of publication.

1. Introduction

In recent years, psychologists have been urged to report measures of the strength of the relationship between their independent and dependent variables (Hays, 1963; Bakan, 1966; Vaughan & Corballis, 1969; Dodd & Schultz, 1973; Dwyer, 1974). For example, one leading journal (*Learning and Motivation*) now requests authors to report a measure of the size of their effects, such as ω^2, a fact that testifies to a growing belief that such measures convey information beyond that conveyed by tests of significance. We suggest, however, that given the present editorial emphasis on statistical significance, estimates of the size of experimental effects are badly inflated, and therefore of questionable value.

Because non-significant results are rarely published, Sterling (1959) argued that the published literature may contain a larger proportion of Type I errors than the stated alpha level. Greenwald (1975), on the other hand, on the basis of a simulation of the publication system concluded that the Type I error rate is probably relatively low. He pointed out that the number of Type I errors depends on both the significance level required for publication and proportion of studies undertaken in which the null hypothesis is actually true. While he discussed many of the consequences of publishing only statistically significant results, he did not consider how this practice influences the accuracy of estimates of effect size. Since many published experiments have relatively low power (Cohen, 1962), those that, by chance, find larger differences between means than actually exist in the population will be much more likely to be published than those that by chance, find an effect smaller than the population effect. The selective pressure of using a criterion of statistical significance, therefore, must result in consistent overestimation of the size of experimental effects in the published literature. Certainly, a consideration of this overestimate with respect to nominal alpha level, sample size and size of actual population difference is necessary in order to evaluate published estimates of effect size.

2. Method

To determine the extent of the bias in reported effect size, we simulated the following experimental situation. Suppose that many psychologists were interested in investigating

whether a difference in IQ exists between two groups; that for each group the population standard deviation was 16; and the IQ difference between population means was either small (4 IQ points), medium (8 IQ points) or large (16 IQ points). The choice of these differences coincides with Cohen's (1962) definition of a small effect as 0·25 standard deviation units, a medium effect as 0·5 standard deviation units and a large effect as 1 standard deviation. Each experiment compared two groups with sample sizes of either 5, 10, 15 or 20 subjects per group. Since the null hypothesis was false for all experiments, a Type I error could never occur. Rather, our interest was in determining the extent of the bias introduced in estimating the size of IQ differences between groups when using only the sample means from those experiments that were statistically significant; and how this overestimation of group differences would vary as a function of sample size, size of population difference and the alpha level.

A computer program was written to generate data sets for 5000 'experiments' for each combination of the three effect sizes with each of the four sample sizes. The data were pseudo-random normal deviates generated by subroutine GAUSS from IBM's (1969) Scientific Subroutine Package. The average difference between treatment groups, the average within-groups standard deviation and the average value of ω^2 were calculated for those experimental results significant at alpha levels of 0·001, 0·005, 0·01, 0·025, 0·05, 0·10, 0·20 and 1·0, as determined by analysis of variance.

3. Results and conclusions

The average differences found between group means for each effect size, sample size and alpha level is shown in Table 1. Clearly the distortion in effect size was quite large. For example, when the true difference between groups of ten subjects was 8 IQ points and the alpha level was set at 0·05, the average size of the reported difference between the two groups was over 18 points. Although using a more stringent alpha level is usually thought of as a conservative measure, the estimated difference between groups using a 1 per cent significance level was more than 20 IQ points for the same experiment. The more conservative the alpha level, the smaller the sample size and the smaller the population difference, the greater the distortion. Even with sample sizes of 20 the misrepresentation of group differences is enormous except when quite large differences actually exist; under this condition an actual 4 IQ point difference was estimated to be more than 11 points at the 5 per cent level and more than 14 points at the 1 per cent level. The values in Table 1 are averages of all significant experimental outcomes, including significant differences between means in the direction opposite to that in the population. If such experiments were not included, the bias would have been even greater.

Also shown in Table 1 are average within-group standard deviations for the various conditions. Although the bias in estimated within-group dispersions was not as large as with estimated differences between group means, smaller sample sizes, more stringent alpha levels and smaller population differences led to lower estimates of the standard deviation. As confidence intervals are based on estimated standard deviations, these results indicate that conditions which increase the overestimation of the size of estimated between-group differences also lead to narrower confidence intervals about these biased estimates.

The average value of ω^2 for the various simulated experiments can be seen in Fig. 1. The ω^2 statistic which represents the estimated ratio of treatment variance to total variance shows the same pattern of distortion found in estimates of the difference between group means. When ω^2 actually equalled 0·0588 (approximately 6 per cent of the

Table 1. Estimated differences between means and estimated standard deviations for experiments that were significant at the 5 per cent and 1 per cent levels as functions of sample size and actual population difference

Significance level	Actual population difference	Estimated difference between means: Sample size 5	10	15	20	Estimated standard deviation [a]: Sample size 5	10	15	20
5 per cent	Small 4	15·43 (279, 53) [b]	15·33 (300, 22)	13·21 (476, 15)	11·56 (611, 20)	12·31	14·23	14·95	15·32
	Medium 8	22·21 (511, 19)	18·28 (831, 3)	14·87 (1313, 2)	13·28 (1706, 0)	12·72	14·73	15·28	15·61
	Large 16	26·50 (1442, 2)	21·19 (2569, 0)	18·43 (3746, 0)	17·14 (4330, 0)	13·76	15·39	15·80	15·90
1 per cent	Small 4	21·46 (65, 5)	18·60 (96, 5)	16·38 (125, 3)	14·24 (57, 3)	9·94	13·31	14·46	14·90
	Medium 8	24·95 (144, 3)	20·66 (337, 1)	17·37 (506, 0)	15·66 (700, 0)	10·42	13·99	14·71	15·23
	Large 16	30·37 (480, 0)	23·69 (1410, 0)	20·42 (2496, 0)	17·97 (3313, 0)	11·98	14·75	15·47	15·70

Note: Each estimate is based upon 5000 computer generated experiments.

[a] The true standard deviation was 16·0.

[b] The numbers in parentheses are the number of occasions upon which a significant difference between means was found in the same direction (first number) and in the opposite direction (second number) to that of the population.

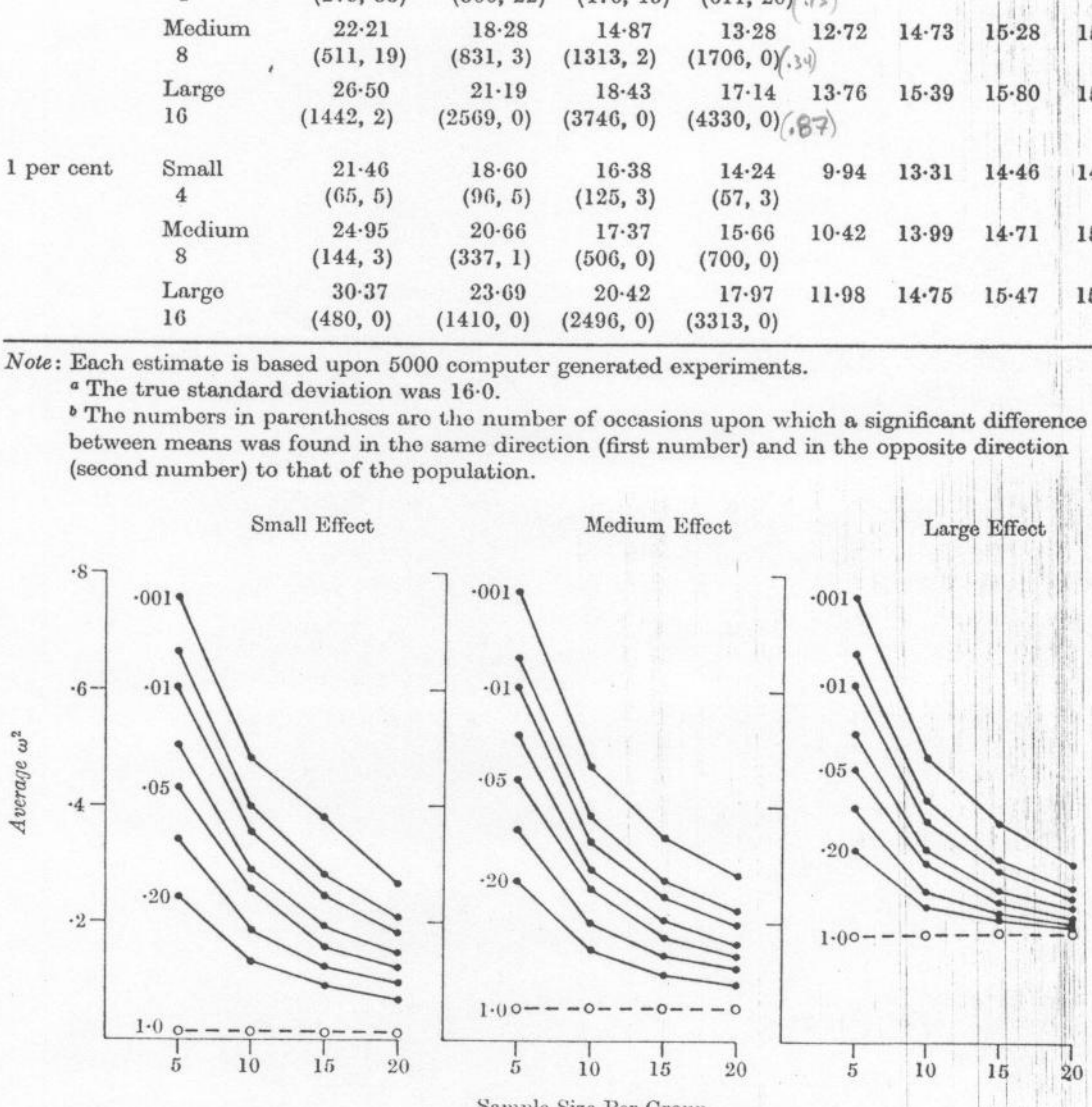

Figure 1. Average ω^2 value for experiments exceeding alpha levels of 0·001, 0·005, 0·01, 0·025, 0·05, 0·10, 0·20 and 1·0 for small (4 IQ points), medium (8 IQ points) and large (16 IQ points) differences between groups as a function of sample size. The population values for ω^2 were 0·0154, 0·0588 and 0·2000 for small, medium and large effect sizes respectively.

total variance was actually attributable to treatment effects) and ten subjects per group were used, the average ω^2 estimated by those experiments significant at the 0·05 level was above 0·25.

An even more interesting finding regarding the ω^2 statistic is that given a fixed group size, published estimates of ω^2 appear to depend mainly on the significance level used by the journal to select articles for publication, and to depend very little on the actual proportion of total variance that would be accounted for by the independent variable in the population. In the present example, the population values of ω^2 are 0·0154, 0·0588 and 0·2000 for groups differing by 4, 8 and 16 IQ points respectively. Considering only experiments significant at the 1 per cent level with ten subjects per group the estimated ω^2 values were 0·354, 0·343 and 0·378 for the 4, 8 and 16 IQ point conditions respectively; at the 5 per cent level the estimates were 0·254, 0·263 and 0·304. There is not even an overlap between the sets of estimates at the two significance levels although the population ω^2 values differ considerably.

The broken lines at the bottom of each panel of Fig. 1, that represent the average ω^2 from all experiments regardless of significance, are of interest in several respects. First, there is a slight bias toward underestimating the actual population values of ω^2. The possibility of such bias was discussed by Vaughan & Corballis (1969), but clearly the bias is not great enough to warrant a proscription against ω^2 on these grounds alone. Second, in order to obtain a reasonable approximation to the population value of ω^2 when averaging across all experiments, it was necessary to include in that average those negative values of ω^2 which occur whenever F ratios are less than unity. Although the concept of a negative proportion of the total variance is meaningless, and one is tempted to report ω^2 equal to zero in such a case, the clear implication is that this should not be done. Setting ω^2 equal to zero when it is negative would make its expected value greater than the population ω^2. Only if negative values are reported, regardless of their apparent lack of meaning, will the actual proportion of variance explained be estimated properly across experiments.

One might think that the overestimation of effect size created by not publishing non-significant results occurs only if several investigators perform the same experiment; however, this is not the case. Any single experiment has a better chance of being published if it overestimates rather then underestimates the strength of the relationship between its independent and dependent variables. The published literature, therefore, must contain a disproportionate number of overestimates.

Clearly, large sample sizes result in less distortion than do small sample sizes. Although the use of large samples cannot eliminate the bias, if one is interested in estimating the size of an effect it is desirable to use as large a sample as possible. However, because time and resources are often limited, insistence on very large sample sizes is not a practical way to reduce the 'publication bias'. Curiously, several authors who have advocated estimating effect sizes have also advocated the use of small samples for reasons other than practical considerations. Bakan (1966) claimed that one can have more confidence in significant results found with small samples than in significant results found with large samples. Hays (1963) cautioned that if one uses too large a sample size one may detect a reliable but trivial effect. It is our contention that if one finds a trivial effect with a large sample, at least one has a fairly good idea of how small the effect actually is. The real danger lies in using small samples, for in that case bias introduced by the significance criterion is greater, thus one is more likely to attribute importance to effects that are trivial in the population.

The present findings indicate that the more conservative the alpha level required by a journal the more distorted are estimates of effect size. Therefore, a journal which rarely

publishes results not significant at the 0·01 level may well contain less accurate estimates of population parameters than those journals that do not require such conservative significance levels. One possibility that has been suggested by Greenwald (1975) is to base publication decisions solely on the introduction and method sections. If a study is theoretically sound and well designed, it should be published regardless of the results. A problem with this approach is that an incompetent experimenter, by introducing extraneous noise, can obscure an effect. Greenwald (1975) discussed this argument, pointing out that sloppy experimenting can also lead to the faulty rejection of the null hypothesis.

A second objection to such a basis for publication decisions is that some studies may be worth while only if they find an effect. For example, a psychologist who develops a new theory may perform an experiment that makes sense only in the context of that theory. If no positive results are found, one might not be interested in either the theory or the experiment. Further, a study which appears uninteresting on the basis of its introduction and method may find unexpected and theoretically important results.

Finally, prejudice against non-significant results occurs at all levels of the publication system. An experimenter who fails to reject the null hypothesis usually will not submit his results for publication (Greenwald, 1975). Although a change in publication criteria would lessen this bias, it is doubtful that it would eliminate it. Despite these possible objections, changing the emphasis from results to method and introduction in deciding which studies to publish is a possibility well worth considering.

Dooling & Danks (1975) also question whether psychology is ready to go beyond tests of significance. Their argument is that measures of the proportion of variance attributable to treatments are not really appropriate for fixed effect designs. The 'importance' of a variable is usually as much a function of the specific levels of the variable chosen as anything else. Only in a random effects design, one rarely used by psychologists, is any useful population quantity estimated. The present findings together with Dooling & Danks' theoretical objections support the conclusion that published estimates of magnitude of effect, expressed in terms of ω^2 or simply as the difference between group means, contain little or no valid information. Although Dwyer (1974) found it odd that significance tests have endured in psychology while measures of the extent of effects have been largely ignored, the present analysis makes it clear that given current publication practices the two approaches cannot coexist. Interest in the magnitude of effects presupposes that psychologists are reasonably sure which variables control a given behaviour, and that enlightened journals exist that will publish the results of all well-designed experiments exploring the influence of those variables.

References

Bakan, D. (1966). The test of significance in psychological research. *Psychol. Bull.* **66**, 432–437.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *J. abnorm. soc. Psychol.* **65**, 145–153.

Dodd, D. H. & Schultz, R. F., Jr (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychol. Bull.* **79**, 391–395.

Dooling, D. J. & Danks, J. H. (1975). Going beyond tests of significance: Is psychology ready? *Bull. Psychon. Soc.* **5**, 15–17.

Dwyer, J. H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychol. Bull.* **81**, 731–737.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychol. Bull.* **82**, 1–20.

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

IBM Scientific Subroutine Package (1969). White Plains, N.Y.: International Business Machines Corp., Manual H 20–0205–03.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. statist. Ass.* **54**, 30–34.

Vaughan, G. M. & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychol. Bull.* **72**, 204–213.

Received 19 April 1977; revised version received 29 March 1978

Requests for reprints should be sent to William P. Dunlap, Department of Psychology, Tulane University, New Orleans, Louisiana 70118, USA.

David M. Lane is now at the Department of Psychology, Rice University, Houston, Texas.