EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1980, 40

CONFIDENCE INTERVALS FOR CORRELATION RATIOS

ALLEN I. FLEISHMAN[1]
Medical Research Division
American Cyanamid Company

It is suggested that for fixed effects experiments, the traditional statistical question is inappropriate. It is suggested that variance ratios—the signal to noise ratio ($\sigma^2_a/\ \sigma^2_e$) and the correlation ratio (σ^2_a/σ^2_t)—are more desirable. Criticisms of these values are explored. Finally confidence intervals and percentiles for the random and fixed effects model's signal to noise and correlation ratio are given. The latter being dependent on the evaluation of a family of non-central F distributions. The latter will also give the confidence interval and percentiles of the multiple correlation coefficient in the regression model. The lower limit of the confidence interval can test the traditional null hypothesis. The upper limit of this interval is suggested to be a test which would allow the investigator to accept the null hypothesis.

IN typical experiments, one can make only one of two conclusions. Either one could conclude that "we have enough evidence to believe that the null hypothesis is false" or that "there currently isn't enough evidence to reject the null hypothesis." As everyone knows, one never accepts the null hypothesis. This is nicely paralleled by reality, as there is seldom a situation where one can reasonably (or theoretically) expect virtually no effect from a treatment. The only exceptions I've ever come across are studies dealing with the validity of ESP. For all other experiments (when talking about mean differences), no one, neither the experimenter nor his staunchest opponent, believes that the

[1] I'd like to thank both Dr. Lloyd Humphreys, for many of his ideas and examples which motivated this problem and Dr. Charles Lewis for his contributions, many insights, and invaluable suggestions on earlier drafts of this manuscript.

Requests for reprints should be sent to Allen Fleishman, Medical Research Division, Lederle Laboratories, Pearl River, N.Y. 10965.

population means are identical to the last decimal point. When working with ANOVA this is equivalent to saying that the non-centrality (n.c.) parameter for effect a (λ^2_a) will invariably be non-zero. As the typical null hypothesis (for the t or F statistic) is a lower one tailed test of $\lambda^2_a = 0$, and as the probability that $\lambda^2_a = 0$ is virtually zero, then the test of this a priori false statement should be considered meaningless. If any readers question the above assertion, let them try to imagine any null hypothesis that is still not rejected after one billion subjects are tested.

The formula for the non-centrality parameter is given by:

$$\lambda^2_a = \Sigma N_j \alpha^2_j / \sigma^2_e \tag{1}$$

where N_j is the number of subjects in group j,
α_j is the population effect of treatment j, and
σ^2_e is the population error variance,

or

$$\lambda^2_a = N_t \, \sigma^2_a / \sigma^2_e \tag{2}$$

where N_t is the total number of observations, and
σ^2_a is the variance due to effect a, given by:

$$\sigma^2_a = \Sigma N_j \, \alpha^2_j / N_t \tag{3}$$

Therefore, the ability to reject the null hypothesis is a function of the number of subjects (N), given a fixed experimental design, the alpha level, and a non-zero non-centrality parameter. (While the power is also dependent on deviations from the basic assumptions of the statistical test, it is assumed here that these assumptions have been approximately met.)

Therefore, I believe that there are two appropriate uses of statistics. The first is discovering a parsimonious and theoretically appealing way to describe the direction and order of the differences among the means. This is usually accomplished by forming interesting contrasts among the means. The second use is to describe the magnitude of the differences between means or the size of the effect.

Various approaches have been used to describe the size of the effect. These include: graphical presentations, simple presentation of the means, and various summary statistics. Two of the more appealing summary statistics are the estimates of the population 'signal to noise ratio'—f^2 (Cohen, 1969, p. 267; Scheffe', 1959, p. 227), and the estimates of the population correlation ratio—η^2 (Hays, 1971, p. 489; Peters and Van Voorhis, 1940, p. 319; Kelly, 1935). The formula for the signal to noise ratio in the population is:

$$f^2 = \sigma^2_a/\sigma^2_e \tag{4}$$

or

$$f^2 = \lambda^2_a/N_t \tag{5}$$

The formula for the correlation ratio in the population is:

$$\eta^2 = \sigma^2_a/\sigma^2_t \tag{6}$$

where σ^2_t is the total variance in the experiment,

or

$$\eta^2 = \sigma^2_a/(\sigma^2_a + \sigma^2_e) \tag{7}$$

or

$$\eta^2 = (1.0 + (f^2)^{-1})^{-1} \tag{8}$$

As can be seen both use estimates of the variance due to the experiment in the numerator, the signal to noise ratio has an estimate of the error variance in its denominator and the correlation ratio has an estimate of the total variance in its denominator.

The population correlation ratio in a one way fixed design is closely related to the non-centrality parameter by:

$$\eta^2 = \lambda^2_a/(\lambda^2_a + N_t) \tag{9}$$

When more than one factor is used (e.g., a three way fixed design with factor a as the relevant factor and b and c other factors), the partial correlation ratio, holding the other factors and interactions constant is given by:

$$\eta^2_{a \cdot bc \cdots} = \lambda^2_a/(\lambda^2_a + N_t) \tag{10}$$

or

$$\eta^2_{a \cdot bc \cdots} = \sigma^2_a/(\sigma^2_a + \sigma^2_e) \tag{11}$$

The intrepation of η^2 has generally been the proportion of variance in the population (or the pseudo-population, see below) due to effect a.

Glass and Hakstian (1969) feel that "the most informative means of reporting the data from a fixed-effects comparative experiment . . . [is to] express mean differences in standard deviation units (p. 413)". In the population, for mean j, this would correspond to looking at α_j/σ_e. If one were to square this quantity, weight it proportional to its frequency, and sum, then one would arrive at the signal to noise ratio, a summary statistic.

$$f^2 = \Sigma N_j \alpha^2_j/N_t \sigma^2_e \tag{12}$$

or

$$f^2 = \sigma^2{}_a/\sigma^2{}_e \tag{13}$$

It is also closely related to the non-centrality parameter

$$f^2 = \lambda^2{}_a/N_t \tag{14}$$

and monotonically related to the (partial) correlation ratio

$$f^2 = \eta^2{}_{a \cdot bc \cdots}/(1 - \eta^2{}_{a \cdot bc \cdots}) \tag{15}$$

It is also interesting to note that when we refer to power tables to see the magnitude of type II error (Cohen, p. 405), the relevant parameters are the degrees of freedom, the alpha level and ϕ^2, where ϕ^2 can be given by

$$\phi^2 = N_j f^2 \tag{16}$$

where N_j is the number of subjects per cell (assuming equal N/cell). Therefore, the (partial) correlation ratio and the signal to noise ratio are functionally related to the power (i.e., the probability of achieving statistical significance).

These two statistics, especially the correlation ratio statistics have come into disfavor lately (Glass and Hakstian, 1969; Carroll and Nordholm, 1975; Keppel, 1973). As most of the comments have been about Hays' omega square for fixed effects ANOVA I shall focus my comments toward it. The basic criticisms are:

1. Bias—It is well known that the ratio of unbiased statistics can be (and usually is) biased. Therefore, while both the numerator and denominator are each separately unbiased estimates, their ratio is itself most likely biased.

In a simulation Carroll and Nordholm (1975), found this bias to be of little practical concern. The largest bias reported for omega square, when sampling proportional to the population, was 0.018. Its bias was small with regard to its standard error (0.191). However, they did report that Kelly's epsilon squared was apparently unbiased.

2. Variability—Carroll and Nordholm (1975) suggested that the large standard error of omega and epsilon square, when N is small mitigates their utility to find a potential large effect. Since these statistics behave very much like correlations, this finding might have been expected. The above standard error (0.191) was found with a total of 15 subjects (from three groups).

The only analytic estimate of the variability of Kelly's epsilon square (Kelly, 1935), assumed both N and ϵ^2 were both large. However any "intrepretation of these values from the application of these formulas (standard error of ϵ^2) requires a knowledge of the form of the

distribution of samples (Peters and Van Voorhis, p. 324)." The only applicable distribution that is recommended (the normal) would only apply when $\eta^2 = 0$. Therefore, Kelly's formula which isn't applicable when epsilon square is small or large may be of questionable utility. What is clearly needed is not a variance estimate but the confidence interval itself. This shall be derived in the next section.

3. Size of Effect—Keppel (1973) has suggested that in some areas one is interested in relatively small effects. For example, in a repeated measurements factorial design, large differences among *S*s are usually ignored, but the existence of a linear x linear interaction may be of great concern. The fact that the variance due to this type of interaction may be small in comparison to the total variance (which includes uninteresting method variance, such as subject differences), may be of little or no concern to the experimenter. As a possible remedy, Keppel (1973) himself makes two recommendations. The first is forming a ratio of comparison variance (e.g., linear x linear interaction) not over total variance, but over some 'interesting' variance (e.g., total interaction variance). The second method he suggests is forming a ratio of the comparison variance (σ^2_a) over comparison plus error variance ($\sigma^2_a/[\sigma^2_a + \sigma^2_e]$). This ratio has been referred to as the partial correlation ratio (Cohen, 1969) and epsilon squared (Humphreys and Fleishman, 1974). A third possibility not mentioned by Keppel is the signal to noise ratio. The latter two possibilities have the added feature that a confidence interval could be generated around them (as described below).

4. Errors in Grouping and Unreliability—Some experiments crudely classify subjects into categories based on continuous variables. For example, an experimenter might classify subject's anxiety level as high, medium, or low (even though none of the subjects in the low group would have exactly the same level of anxiety). Still further confounding this problem might be errors in categorization or non-uniformity of treatments within category.

It was this type of problem which caused Fisher (1938, p. 264) to reject the correlation ratio. The solution to these problems was first suggested by Peters and Van Voorhis. They treated ϵ^2 as they would a regular correlation coefficient. That is, they applied the formula for disattenuation. Specifically the reliability estimate they used (p. 323) was based on errors in grouping (by both the independent and dependent variable). Needless to say if one wished to estimate the correlation ratio free of any type of errors of measurement (categorization or inequality of treatment) one could divide the correlation ratio by the squared reliabilities of the independent and dependent variable.

5. Arbitrariness of η^2—It is argued by Glass and Hakstian (1969)

that estimates of the correlation ratio are a function of the choice of categories, the range, and the complexity of the independent variables. They are totally correct in this regard. If one compared two treatments which are very similar to one another (all other things being constant) η^2, the power (related to η^2), and hence the F with its probability would be low in comparison to an experiment where the treatments were widely different. It would be quite consistent for ϵ^2 to be .15 in one case and .75 in a second. They also point out that different operational definitions could invariably produce the above results. For example, one may be interested in the effects of different methods of teaching. In one experiment one could compare the effects of white verses yellow chalk. In a second experiment, the effects of lecture only verses reading only could be compared. These two studies should yield totally different results (in terms of significance, mean differences and η^2).

Up to this point there can be little argument. Statements as to the effects of manipulating any conceptual independent variable, when one is considering the entire range of the variable as seen in the real world, should be done in the framework of a random model, as the vast majority of effects of concern to the psychologist are truly random effects (however there are many exceptions [e.g., sex]). One would have to sample randomly from the true population of treatments to get the true estimate of the proportion of variance explained by the effect.

This does not imply that all fixed effects ANOVAs and associated statistics, including the F, are invalid. When an experimenter generates an experimental design s/he is actually creating a pseudo-population. Results from this pseudo-population may produce results exactly opposite in direction to that which appears in the real world or any one else's pseudo-population, for example, crossovers (Humphreys and Fleishman, 1974; Keren and Lewis, 1976). Nevertheless, if one were to fix the values of the independent variable and then report findings conditional on only those levels selected, this is perfectly valid methodologically.

There seem to be two types of psychologists. One describes the actual world as parsimoniously as possible. The second asks questions like 'what if the world were changed like this . . .'. While these two roles overlap, the former has been the province of the correlational psychologist, the latter the experimentalist. Given all the necessary conditional statements to describe the pseudo-universe created, there is no reason why the proportion of variance explained has less meaning in the experimenter's world than in the actual world. One needn't have to wait for a possible treatment to be implemented in the real

world on a mass level before assessing its effects. By creating a control and experimental group, one can predict the effects of the treatment (including mean differences and assorted statistics—e.g., η^2, F).

Confidence Intervals

As stated above, the best information we have is an estimate (although possibly biased) of the population correlation ratio. No useful information exists for the standard error of this estimate or its confidence interval. Information about the confidence interval for η^2 is primarily useful for the direct interpretations that can be made for its upper and lower limits.

As stated earlier, if the lower limit for λ^2 includes zero, the traditional H_0 would not be accepted. This is equivalent (as by Equations 9 and 10) to a test of the lower limit of the correlation ratio. That is, if the confidence interval for η^2 includes (excludes) zero then one would accept (reject) the traditional null hypothesis. One could easily liberalize this test by testing if the correlation ratio could take on another small value. For example, if the interval has a range of 0.03 to 0.09, then the traditional null hypothesis would be rejected and one can conclude that the experimental manipulation had some effect and could predict more than 3% (therefore, more than 0% and more than 1%) of the variance.

The upper limit of the confidence interval can be used to *accept* a form of the null hypothesis. If a discipline can state a proportion of predicted variance which would be considered 'trivial', and if the upper limit is less than this value, then one could conclude that the manipulation had no practical effect, i.e., it isn't *psychologically* significant. Returning to the above example, if the discipline felt that an experimental manipulation had to predict at least 10% of the variance to be considered 'non-trivial', then since the above upper limit (0.09) of the confidence interval was below 0.10 the results could be thought of as psychologically insignificant. Therefore, the experimental results were statistically significant but practically useless. An experimental manipulation which yielded an interval from 0.00 to 0.60 can be thought of as statistically insignificant, but possibly of great importance. While this manipulation may be useless (i.e., $\eta^2 = 0$), it might also be quite useful (i.e., $\eta^2 = 0.60$). Further research into this area (with more subjects and greater experimental control) may be called for.

Random Effects Model: Scheffe' (1959, p. 225–231) has derived a confidence interval for the case where the experimenter randomly selected the treatment levels from a population of (normally distributed)

effects and the experimenter has control over which treatment groups the subjects belong.

Under these assumptions, the sum of squares between treatments (even when the treatment means in the population are unequal) is proportional to a Chi Square random variable. The distribution of the signal to noise ratio is related to a central F distribution. Furthermore, none of the parameters of the F distribution are dependent on the value of the non-centrality parameter (or ϕ^2 or f^2). The central confidence interval for f^2 is given by:

$$[(F_a/F_U) - 1.0]/J < f^2 < [(F_a/F_L) - 1.0]/J \tag{17}$$

where J is the number of treatment levels,
F_a is the sample F ratio for effect a,
F_L is the critical[2] lower point $(\alpha/2)$ of the F distribution, with d.f. ν_n and ν_d, and
F_U is the critical[2] upper point $(1 - \alpha/2)$ of the F distribution, with d.f. ν_n and ν_d.

The confidence interval for the correlation ratio is obtained by use of Eqn. 8, and is given by,

$$(1.0 + L^{-1})^{-1} < \eta^2 < (1.0 + R^{-1})^{-1} \tag{18}$$

where L and R are respectively, the left and right hand members of Equation 17.

The confidence limits, for both f^2 and η^2, can take on negative values. Although the true value for either f^2 or η^2 must be non-negative, Scheffe' recommends not changing negative values (although they may be illogical) to zero. Scheffe' made his recommendation (p. 229–231) on "intuitive grounds," primarily to convey the greater certainty that the true parameter may be zero. The amount below zero should be thought of only as a non-mathematical aid. Negative variance estimates are, of course, meaningless. However allowing one or both (if F_a were sufficiently small) of the limits to go from inconsistent negative values to zero will not affect the probability statement. In addition, the replacement by zero will yield a shorter confidence interval.

Fixed Effects Model: When the population treatment effect has an unknown distribution, but one has observed all levels of the treatment (i.e., one is not sampling treatment levels), and the experimenter has control over which treatment group any subject enters, one can't find the confidence interval of f^2 or η^2 by use of a simple approach.

In order to find the confidence interval, one must take a more general approach (Mood and Graybill, p. 256–260; Kendall and Stuart, p. 103–105). This approach may be most easily understood by Figure 1. Let F be any observed sample F statistic. Let λ^2 be the true population

n.c. parameter. In this general approach all possible sample F ratios shall be postulated. Furthermore the population n.c. parameter shall be allowed to assume any value. For any single value of the population n.c. parameter (e.g., λ^2_0) a confidence interval may be obtained. One such interval is indicated by the hatch marks on the vertical line above λ^2_0. All such confidence intervals, one set of points for every possible population n.c. parameter, may be obtained and produce what is known as a confidence belt (Kendall and Stuart, p. 100). This

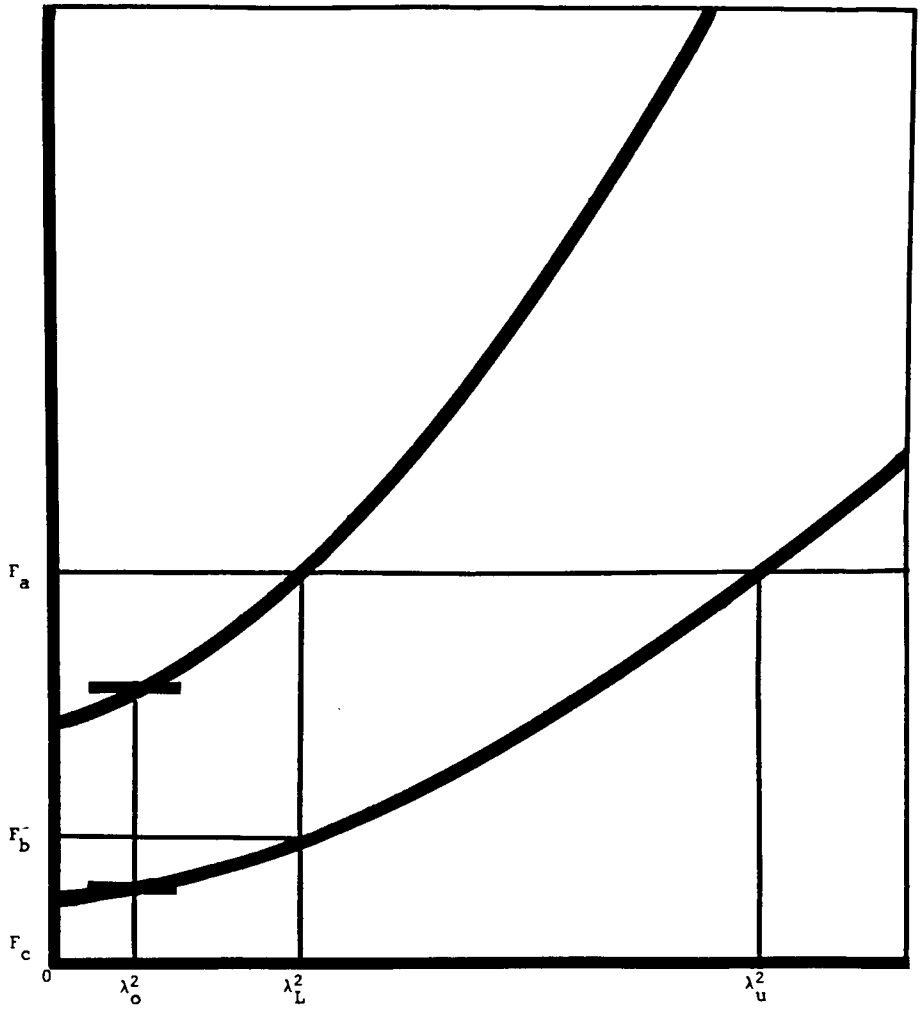

Figure 1. Confidence belt for the non-central F distribution. X axis: Non-centrality parameter. Y axis: Observed F statistics.

belt (made up of two curves) is presented in Figure 1 by the diagonal curves. Therefore, the points within the confidence belt consist of all those F statistics seen $(1 - \alpha)$ proportion of the time for any possible value of λ^2. Although the belt was created by vertical entries, in a given experiment one obtains a particular value of F (e.g., F_a). One may then find all those population values which could reasonably produce this sample value by entering this figure horizontally and discovering that given F_a, ${\lambda^2}_L$ cuts off the lower limit and ${\lambda^2}_U$, the upper limit. In other words, F_a is in the acceptance region of every possible value of λ^2 between ${\lambda^2}_L$ and ${\lambda^2}_U$. F_a falls in the acceptance region of ${\lambda^2}_o$ if and only if ${\lambda^2}_o$ falls in the confidence interval based on F_a. Once one obtains ${\lambda^2}_L$ (or ${\lambda^2}_U$), then the upper and lower limits of f^2 may be obtained via Equation 5.

Actually one need not find the complete confidence belt. For the lower limit, one needs to search for that value of λ^2, ${\lambda^2}_L$, given the obtained F ratio, F_a, and the experimental degrees of freedom (ν_n and ν_d) such that:

$$\alpha/2 = \text{Prob}\,(F < F_a \mid {\lambda^2}_L, \nu_n \text{ and } \nu_d) \tag{19}$$

where the above probability statement is the cumulative probability that a sample n.c. F will be less than a given F, given a population non-centrality parameter, λ^2, and the appropriate *d.f.*

The corresponding upper limit, ${\lambda^2}_U$, may be found by searching for that value of λ^2, ${\lambda^2}_U$, such that:

$$1 - \alpha/2 = \text{Prob}\,(F < F_a \mid {\lambda^2}_U, \nu_n \text{ and } \nu_d) \tag{20}$$

Both of the above probabilities can be found by use of the n.c. F distribution. This is because the SS_{effect} will now follow a n.c. Chi Square distribution. The ratio of the MS_{effect} over the MS_{error} will be a (singly) *n.c. F.*

If the F ratio (e.g., F_b) is in the acceptance region for the traditional[2] null hypothesis (H_o:$\lambda^2 = 0$), then there is no lower limit, ${\lambda^2}_L$, possible as ${\lambda^2}_L = 0$ is the smallest value the n.c. parameter can take.

If a point is unobtainable (this is analogous to the use of a negative limit for the random model) one may say it is indistinguishable from zero. Given a particularly small F statistic (e.g., F_c), it is conceivable that the upper critical limit will also be impossible to obtain. In this case the interval will be from 0.0 to 0.0. The length of the confidence interval in this case should not be used as an indicant of the accuracy

[2] These critical points on the F distribution will not correspond to the typical critical F value, as the latter usually refers to an upper *one* tailed test (i.e., a point corresponding to $[1 - \alpha]$ on the F distribution). The traditional null hypothesis as used in this paper will be a two tailed test.

of the estimation procedure (in contrast to typical intervals where the interval length is proportional to the standard error).

Regardless of what the true value of λ^2 is, the probability that the interval constructed by the above technique will include λ^2 is $1-\alpha$.

By the appropriate choice of α, confidence intervals of various lengths can be obtained. When the length is zero, $\alpha = 0.50$, the median, $\lambda^2_{mdn.}$, can be found. Over repeated samples, 50% of the time $\lambda^2_{mdn.}$ will be greater than the true λ^2. In other words, the median of this sampling distribution will be the true parameter. As this distribution is most likely highly skewed, the median is to be preferred to an estimate of the mean. However the unbiased estimate of the mean and the mean square error (MSE) of f^2 may be easily found by:

$$E(f^2) = [F_a\, \nu_n(\nu_d - 2)/\nu_d - \nu_n]/N_t \tag{21}$$

and

$$MSE(f^2) = 2[(\nu_n + N_t E(f^2))^2 + (\nu_n + 2N_t E(f^2))(\nu_d - 2)]/\, N^2_t(\nu_d - 4) \tag{22}$$

The above unbiased estimate of the mean and MSE was found by a linear transformation on the mean and variance of the sampling distribution of λ^2_a (Johnson and Kotz).

The appropriate percentile points of the correlation ratio are found by use of Equation 10. While a monotone transformation doesn't affect the estimation of probabilities, it will affect means and variances. Therefore, one can obtain the confidence interval for η^2, as well as its median, but not its mean or variance.

By use of this approach, and the identity of ANOVA and regression via the general linear model, the confidence interval and median of the population correlation coefficient in the regression model may be obtained.

In order to find the above points one must be able to evaluate the probability integral of the n.c. F distribution and iteratively search for the limits. A program was written by this author to find Prob ($F < F_a$ given ν_n, ν_d, and λ^2_a). The program uses a form of the infinite series needed for this integral given by Venables (1975, p. 410, Eqn. 12). This equation was chosen as it appears to utilize simpler operations (from a computer's point of view) than analogous formulea (Johnson and Kotz, 1970, Eqn. 12 and 13). A subroutine using a halving procedure[3], was found to be the most efficient search algorithm.

Example: Venables gives an illustration where $\nu_n = 4$, $\nu_d = 50$ and $F_a = 11.2213$.

[3] This single parameter search subroutine was made available by Ronald L. Hinkle of the University of Illinois at Champaign-Urbana.

	Percentile Points		
	0.05	0.50	0.95
λ^2_a	19.381	41.373	71.553
f^2	0.352	0.752	1.301
η^2	0.260	0.429	0.565

$E(f^2) = 0.711$, $MSE(f^2) = 0.083$ and Hays omega square = 0.426.

A program to evaluate the fixed effects confidence interval is available from the author for either CDC or IBM computers.

REFERENCES

Carroll, R. M. and Nordholm, L. A. Sampling characteristics of Kelly's ϵ^2 and Hays' $\hat{\omega}^2$. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENTS, 1975, 35, 541–554.

Cohen, J., *Statistical power for the behavioral sciences.* New York: Academic Press, 1969.

Fisher, R. A. *Statistical methods for research workers,* (7th ed.). Edinburgh: Oliver and Boyd, 1938.

Glass, G. V and 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 for the social sciences.* New York: Holt, Rinehart and Winston Inc., 1973.

Humphreys, L. G. and Fleishman, A. I. Pseudo orthogonal and other analysis of variance designs involving individual difference variables. *Journal of Educational Psychology,* 1974, 66, 464–472.

Johnson, N. I. and Kotz, S. *Continuous Distributions—2,* Boston: Houghton Mifflin Co., 1970.

Kelly, T. L. An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences,* 1935, 21, 554–559.

Kendall, M. G. and Stuart, A. *The advanced theory of statistics,* 2, (2nd ed.) London: Charles Grifflin and Co. Ltd. 1967.

Keren, G. and Lewis, C. Nonorthogonal designs: sample verses population. *Psychological Bulletin,* 1976, 83, 817–826.

Mood, A. M. and Graybill, F. A. *Introduction to the theory of statistics,* (2nd ed.) New York: McGraw Hill Book Co., Inc., 1963.

Peters, C. C. and Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw Hill Book Co., Inc., 1940.

Scheffe', H. *The analysis of variance.* New York: John Wiley and Sons, Inc., 1959.

Venables, W. Calculation of confidence intervals for non-centrality parameters. *Journal of the Royal Statistical Society,* Series B, 37, 1975, 406–412.