Psychological Bulletin
1992, Vol. 111, No. 1, 188-194

0033-2909/92/$3.00

Testing the Null Hypothesis in Meta-Analysis: A Comparison of Combined Probability and Confidence Interval Procedures

Larry V. Hedges
Department of Education
University of Chicago

Harris Cooper
University of Missouri

Brad J. Bushman
Iowa State University

Combined significance tests (combined p values) and tests of the weighted mean effect size are both used to combine information across studies in meta-analysis. This article compares a combined significance test (the Stouffer test) with a test based on the weighted mean effect size as tests of the same null hypothesis. The tests are compared analytically in the case in which the within-group variances are known and compared through large-sample theory in the more usual case in which the variances are unknown. Generalizations suggested are then explored through a simulation study. This work demonstrates that the test based on the average effect size is usually more powerful than the Stouffer test unless there is a substantial negative correlation between within-study sample size and effect size. Thus the test based on the average effect size is generally preferable, and there is little reason to also calculate the Stouffer test.

In the past two decades, there has been increasing interest in the use of systematic procedures for combining evidence in literature reviews. One aspect of these procedures has been the use of quantitative methods for combining results of statistical analyses across studies or meta-analysis (Cooper, 1989). Although statistical methods for combining the results of independent studies have a history dating to at least the 1930s (e.g., Cochran, 1937; Fisher, 1932, p. 99), their widespread application to research in the social and behavioral sciences (and the term *meta-analysis* itself) is relatively new (see Glass, McGaw, & Smith, 1981).

There are two basic approaches to combining evidence across studies in meta-analysis. One approach involves testing the statistical significance of combined results of the collection of studies. That is, testing whether the observed collection of results could have arisen by chance if the null hypothesis were true in every study. The second approach involves estimating a combined (average) treatment effect. A confidence interval or significance test is often used to determine whether the combined treatment effect is reliably different from zero (Hedges & Olkin, 1985).

Meta-analysts who rely on the combined significance testing approaches emphasize (or should emphasize) testing of the null hypothesis that the (treatment) effect is zero in all studies. Meta-analysts who rely on the combined estimation approach emphasize the characterization of the magnitude of the combined effect. They may, however, compute a confidence interval for that combined effect or compute a test that the combined effect is different from zero. Some meta-analysts use both combined significance tests and estimate and test the combined effect. Although the two approaches use different information from each study (combined significance tests use p values and combined estimation procedures use measures of effect size), the methods are clearly related (see Becker, 1987).

The purpose of this article is to clarify the general relationship between these two approaches to meta-analysis. We do so by comparing the properties of the statistical tests involved in the most widely used combined significance test (the Stouffer or inverse normal procedure; see Rosenthal, 1984) to those of the most widely used test of the combined effect size (the test of the weighted mean effect size; see Hedges & Olkin, 1985). We show that both procedures can be used to test the same hypothesis. Then we show that the two tests give similar results when applied to the same data in the sense that the power of the tests is quite similar.

It might seem surprising that a combined significance test, which is not explicitly weighted, gives results that are quite similar to those of a test based on weighted combinations of effect sizes. The reason is that the unweighted combined significance test is actually weighted indirectly by sample size through the p value. The self-weighting of the combined test procedure is most obvious in the case of combining tests with known variances. We show that in this case the Stouffer test statistic can actually be written as a sample-size-weighted mean of sample effect sizes.

We begin by providing the statistical model, notation, and the definition of the two tests. Then we derive the properties of the two tests in the situation in which the variance within treat-

---

The simulation study reported in this article was submitted in partial fulfillment of Brad J. Bushman's requirements for the degree of master of arts in statistics at the University of Missouri.

We express sincere appreciation to Paul Speckman for his helpful comments on a draft of this article, to Wayne Churchill for his help with the FORTRAN simulation program, and to Margie Gurwit for her help with the Statistical Analysis System program.

Correspondence concerning this article should be addressed to Larry V. Hedges, Department of Education, University of Chicago, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

ment groups is known. This provides an approximation to the more realistic case in which variance is unknown and must be estimated. Next we present asymptotic (large sample) results where the variance is unknown. Finally we present the results of a simulation study to support the generalizations suggested by the analytic results. The simulations systematically vary five factors: (a) the number of studies in the meta-analysis, (b) the population value of effect size underlying the set of studies, (c) the variance of the sample effect sizes, (d) the sample sizes contained in individual studies, and (e) the correlation between study sample sizes and population effect sizes.

Model and Notation

Suppose that the data arise from a series of k independent studies, each of which compares a treatment or experimental group (E) with a control group (C). Let Y_{ij}^{E} and Y_{ij}^{C} denote the jth observations in the experimental and control groups of the ith study, and let n_i^{E} and n_i^{C} be the experimental and control group sample sizes. Suppose also that the assumptions for the validity of the two-sample t test are met in each study. That is, the observations in the experimental and control groups of the ith study are independently normally distributed with means μ_i^{E} and μ_i^{C}, respectively, and common variance σ_i^2. Define the (population) effect size in the ith study as the standardized mean difference

$$\delta_i = \frac{\mu_i^{E} - \mu_i^{C}}{\sigma_i} \tag{1}$$

and define the sample estimate of effect size (the sample effect size) in the ith study as

$$d_i = \frac{\bar{Y}_i^{E} - \bar{Y}_i^{C}}{s_i}, \tag{2}$$

where $\bar{Y}_i^{E}$ and $\bar{Y}_i^{C}$ are the sample means in the experimental and control groups, respectively, and s_i is the pooled within-group sample standard deviation. The two-sample t statistic for testing the significance of mean differences in the ith study can be written as

$$t_i = \sqrt{\tilde{n}_i} d_i, \tag{3}$$

where $\tilde{n}_i$ is the harmonic mean of n_i^{E} and n_i^{C} given by

$$\tilde{n}_i = \frac{n_i^{E} n_i^{C}}{n_i^{E} + n_i^{C}}.$$

Let p_i be the one-tailed p value for the ith study derived from t_i. Thus p_i is the probability of obtaining a t statistic larger than t_i in the ith study when the null hypothesis is true.

Combined Significance Tests in Meta-Analysis

Rosenthal (1978, 1984) described 7 methods for combining the probabilities of independent studies. Hedges and Olkin (1985) described 4 additional methods and discussed the statistical properties of all 11 methods. Rosenthal (1978) argued that the method of adding zs is the most serviceable because of its simplicity and general applicability. Today this method is probably most often used by meta-analysts and is compared in this article to the test of the mean effect size. The method was first developed by Stouffer, Suchman, DeVinney, Star, and Williams (1949; p. 45) and is often referred to as the *Stouffer method*. The Stouffer method involves (a) converting one-tailed p levels to their associated z scores, (b) retaining the direction of each study's outcome by attaching a positive or negative sign to the z score depending on whether the directional hypothesis was supported, (c) summing the z scores, (d) dividing by the square root of the number of p levels, and (e) referring this number back to a standard normal distribution table to obtain a combined significance level (one-sided).

Thus, the Stouffer method could be described symbolically by saying that to test the joint null hypothesis,

$$H_0: \delta_1 = \cdots = \delta_k = 0,$$

compute the statistic

$$Z_S = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} z(p_i)$$

where $z(p_i) = -\Phi^{-1}(p_i)$ is the z score corresponding to p_i, the one-tailed p value associated with t_i, the t statistic in study i. We reject H_0 at significance level α if Z_S exceeds the $100\alpha\%$ one-tailed critical value of the standard normal distribution.

Effect Size Estimation in Meta-Analysis

Two measures of effect dominate the meta-analytic literature. When the primary studies in question compare two groups, either through treatment versus control comparisons or through single-degree-of-freedom contrasts, the effect size is expressed as some form of standardized difference between the group means, often called a d index (Cohen, 1977). When two continuous variables are related, the product–moment correlation coefficient, or r index, is most often used. We restrict our attention to the d index in this article, though similar results hold for the r index.

The (weighted) mean effect size in meta-analysis is calculated by averaging the individual effects after each has been weighted by the inverse of its variance (Hedges & Olkin, 1985, pp. 110–113). The standard error of the weighted mean effect size is the square root of the reciprocal of the sum of the weights. The test that the mean effect differs from zero consists of (a) computing the weighted mean effect size, (b) dividing it by its standard error, and (c) referring the ratio to a standard normal table to obtain a significance level.

The one-tailed test that the average effect size is greater than zero can be described symbolically by saying that to test H_0, compute the test statistic

$$Z_{\mathrm{LCL}} = \frac{d.}{S(d.)},$$

where $d.$ is the weighted mean effect size and $S(d.)$ is its standard error. More specifically,

$$Z_{LCL} = \frac{\sum_{i=1}^{k} w_i d_i}{\sqrt{\sum_{i=1}^{k} w_i}},$$

where the weight w_i, given by

$$w_i = \frac{1}{\tilde{n}_i} + \frac{d_i^2}{2(n_i^E + n_i^C)},$$

is the reciprocal of the estimated sampling variance of d_i (Hedges & Olkin, 1985; p. 115).

We reject H_0 at significance level α if Z_{LCL} exceeds the $100\alpha\%$ one-tailed critical value of the standard normal distribution. This test is algebraically equivalent to, but more direct than, computing a confidence interval about the average effect size and rejecting the null hypothesis that the effect size is zero if the confidence interval does not contain zero. Because the one-sided test on the mean effect size is mathematically equivalent to the test that is based on whether the lower confidence limit exceeds zero, we refer to the test on the mean effect size as the lower confidence limit (LCL) test.

Exact Theory When the Within-Group Variance Is Known

It is useful to compare the Stouffer test with the LCL test in the case in which the within-group standard deviation σ_i is known in each study. Exact theory can be obtained in this case, and it provides an approximation to what might be expected in the more complex (but more realistic) case when the standard deviations are not known. When the standard deviations are known, the (optimal) statistical test for treatment effects (that is, to test whether $\mu^E = \mu^C$) is the z test. The z statistic in the i th study is

$$z_i = \sqrt{\tilde{n}_i}\left[\frac{\bar{Y}_i^E - \bar{Y}_i^C}{\sigma_i}\right] = \sqrt{\tilde{n}_i} d_i \tag{5}$$

where

$$d_i = \frac{\bar{Y}_i^E - \bar{Y}_i^C}{\sigma_i} \tag{6}$$

is an effect size estimate computed using the population standard deviation σ_i, which is assumed to be known.

Both the LCL test and the Stouffer test yield statistics (Z_{LCL} and Z_S, respectively) that have standard normal distributions when the joint null hypothesis

$$H_0\colon \delta_1 = \cdots = \delta_k = 0$$

is true. Thus, both the LCL test and the Stouffer test involve comparing the values of their respective z statistics to critical values obtained from the standard normal distribution. The decision rules for one-sided tests at the $100\alpha\%$ significance level are for LCL, reject H_0 if $Z_{LCL} > C_\alpha$, and for Stouffer, reject H_0 if $Z_S > C_\alpha$ where C_α is the $100\alpha\%$ critical value of the standard normal distribution.

The LCL test statistic, Z_{LCL}, can be written as

$$Z_{LCL} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left[\frac{\tilde{n}_i}{\bar{n}}\right]^{1/2} \sqrt{\tilde{n}_i} d_i \tag{7}$$

where $\bar{n}$ is the average of $\tilde{n}_1, \ldots, \tilde{n}_k$. The Stouffer test statistic, Z_S, can be written as

$$Z_S = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \sqrt{\tilde{n}_i} d_i. \tag{8}$$

Equation 8 reveals that the Stouffer test statistic, although not explicitly weighted, is actually equivalent to a sample size weighted mean of effect size estimates. The "weighting" is a consequence of the fact that the test statistic used to generate the p values depends on sample size as well as effect size.

To evaluate the power of these two tests, we need to know the probability that each will reject the null hypothesis when H_0 is false. This probability can be obtained through the sampling distribution of the test statistic when H_0 is false. The sampling distributions of Z_{LCL} and Z_S are both normally distributed with a variance of one, but the expression for the means can be different. In particular,

$$Z_{LCL} \sim N(\delta_{LCL}, 1), \quad Z_S \sim N(\delta_s, 1),$$

where

$$\delta_{LCL} = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \left[\frac{\tilde{n}_i}{\bar{n}}\right]^{1/2} \sqrt{\tilde{n}_i}\delta_i, \quad \delta_s = \frac{1}{\sqrt{k}} \sum_{i=1}^{k} \sqrt{\tilde{n}_i}\delta_i. \tag{9}$$

An exact expression for the power of these two tests at significance level α is

$$\text{power(Stouffer test)} = 1 - \Phi(C_\alpha - \delta_s) \tag{10}$$

and

$$\text{power(LCL test)} = 1 - \Phi(C_\alpha - \delta_{LCL}) \tag{11}$$

where $\Phi(x)$ is the standard normal cumulative distribution function.

Formulas 9, 10, and 11 imply that the power of these tests depends on $\delta_1, \cdots, \delta_k$, and on the sample sizes. Because δ_{LCL} and δ_s combine $\delta_1, \cdots, \delta_k$ in different ways, neither of these tests is the most powerful in all situations. However, a few generalizations are possible. When the sample sizes in all of the studies are equal (that is, when $\tilde{n}_1 = \cdots = \tilde{n}_k$), the tests yield the same tests statistics and, therefore, have identical power. If the sample sizes are unequal but all of the studies have the same effect size (that is, if $\delta_1, = \cdots = \delta_k = \delta$), then $\delta_{LCL} > \delta_s$, and the LCL test is more powerful than the Stouffer test. If both sample size and population effect sizes are unequal, then either the LCL test or the Stouffer test may be more powerful. If larger values of $\tilde{n}_i$ are associated with larger values of δ_i (that is, if δ values are positively correlated with $\tilde{n}$ values), then $\delta_{LCL} > \delta_s$, and the LCL test is more powerful than the Stouffer test. If δ values are negatively correlated with $\tilde{n}$ values, then δ_s can be larger than δ_{LCL}, and the Stouffer test can be more powerful than the LCL test.

Asymptotic Theory When the Within-Group Variance Is Unknown

Examination of the case in which the within-group variance σ_i^2 is known in each study provides useful insight into the behavior of the two tests, but it provides only a rough approximation of their behavior when the variances are unknown. For example, although the individual t statistics in each study tend to z statistics in large samples (that is, in large samples the t statistic tends to behave essentially as if the variance σ_i^2 were known), the large-sample distribution of $z(p_i) = -\Phi^{-1}(p_i)$ is not the same when p_i is based on a z statistic as it is when p_i is based on a t statistic (see Lambert, 1978; Lambert & Hall, 1982, 1983). Because the Stouffer test is computed from $z(p_i)$ values, it therefore does not have precisely the same behavior, even in large samples, when the variance is known as when it is unknown. The large-sample distribution of Z_{LCL} also differs when the variance is unknown. Thus it is necessary to compare the properties of the two tests when the variances are assumed to be unknown.

Asymptotic Theory for the Stouffer Test

Becker (1985) has used results from the asymptotic efficiency of test statistics to obtain the asymptotic distribution of $z(p)$. She showed that if $n_i = n_i^E + n_i^C$, $\pi_i^E = n_i^E/n_i$, $\pi_i^C = n_i^C/n_i$ and π_i^E and π_i^C remain fixed as $n_i \to \infty$, then the asymptotic distribution of $z(p_i)$ when $n_i \to \infty$ and $\delta > 0$ is given by

$$[z(p_i) - \sqrt{n_i \log(1 + \pi_i^C \pi_i^E \delta_i^2)}] \sim N(0, \eta_i^2), \qquad (12)$$

where

$$\eta_i^2 = \frac{\pi_i^E \pi_i^C \delta_i^2 (1 + \pi_i^E \pi_i^C \delta_i^2/2)}{(1 + \pi_i^E \pi_i^C \delta_i^2)^2 \log(1 + \pi_i^E \pi_i^C \delta_i^2)}. \qquad (13)$$

This asymptotic distribution shows the limitations of the approximation of $z(p_i)$ (where p_i is computed from a t statistic) by $\sqrt{\tilde{n}_i} d_i$ (the value assuming σ_i is known). Neither the mean ($\sqrt{\tilde{n}_i}\delta_i$) nor the variance (1) of the $z(p_i)$ when σ_i is known is the same as that of the limiting distribution of $z(p_i)$ when σ_i is unknown. Expanding the logarithm in Equation 12 through a Taylor series, we found that when δ_i is small, the mean of $z(p_i)$ is approximately

$$\sqrt{n_i \log(1 + \pi_i^E \pi_i^C \delta_i^2)} \approx \sqrt{\tilde{n}_i}\delta_i,$$

and using L'Hospital's rule, the limit of the variance η_i^2 as $\delta_i \to 0$ is 1. Thus the limiting distribution of $z(p_i)$ corresponds to the distribution with σ_i known only for small values of δ_i.

Because the Stouffer statistic Z_s is a linear combination of $z(p_1), \ldots, z(p_k)$, the asymptotic distribution of $z(p_i)$ implies the asymptotic distribution of Z_S. If $N = \sum_{i=1}^k n_i$ and $n_1/N, \ldots, n_k/N$ remain fixed as $N \to \infty$, then the large-sample approximation that is based on the asymptotic distribution of Z_S is given by

$$Z_S \sim N(\mu_s, \eta_s^2), \qquad (14)$$

where

$$\mu_s = \sum_{i=1}^k \sqrt{n_i \log(1 + \pi_i^C \pi_i^E \delta_i^2)}/k, \qquad (15)$$

and

$$\eta_s^2 = \sum_{i=1}^k \eta_i^2/k, \qquad (16)$$

and η_i^2 is given by Equation 13. The power of the Stouffer test for significance level α is just the probability that Z_S is greater than the critical value C_α. Consequently, the power of the Stouffer test computed from the large-sample approximation (14) is just

$$\text{power (Stouffer test)} = 1 - \Phi[(C_\alpha - \mu_s)/\eta_s], \qquad (17)$$

where $\Phi(x)$ is the standard normal cumulative distribution function, μ_s is given by Equation 15, and η_S^2 is given by Equation 16.

Asymptotic Theory for the LCL Test

The asymptotic distribution of the weighted mean effect sizes was given by Hedges (1982) for the case of effect sizes computed under the assumption that the within-group variances are unknown in each study. His results imply that if $n_i = n_i^E + n_i^C$, $N = \sum_{i=1}^k n_i$, and $n_i^E/N, \ldots, n_k^E/N$ remain fixed as $N \to \infty$, then the large-sample approximation to the distribution of Z_{LCL} is

$$Z_{LCL} \sim N(\delta_{LCL}, 1), \qquad (18)$$

where

$$\delta_{LCL} = \sum_{i=1}^k \omega_i \delta_i / [\sum_{i=1}^k \omega_i], \qquad (19)$$

and

$$\omega_i = \frac{1}{\tilde{n}_i} + \frac{\delta_i^2}{2(n_i^E + n_i^C)}. \qquad (20)$$

Hedges (1982) studied the accuracy of this large-sample approximation and found it to be quite accurate for a wide range of sample sizes and effect sizes.

The power of the LCL test for significance level α is the probability that Z_{LCL} exceeds the critical value C_α. Consequently, the power of the LCL test computed from the large-sample approximation is

$$\text{power (LCL test)} = 1 - \Phi(C_\alpha - \delta_{LCL}), \qquad (21)$$

where $\Phi(x)$ is the standard normal cumulative distribution function and δ_{LCL} is given by Equation 19.

Comparing the Two Tests

The algebraic forms of Expressions 17 and 21 for the power of the two sets do not lend themselves to easy comparisons. However, when $\delta_1, \ldots, \delta_k$ are all small, then Equation 17 reduces

approximately to Equation 10, and Equation 21 reduces approximately to Equation 11. Hence for small effect sizes and large sample sizes, we expect the qualitative generalizations derived in the case in which the variance is known to apply to the case in which the variance is unknown. Numerical computations for nonnegligable values of δ_i are also consistent with the expectation that the LCL and Stouffer tests will have approximately the same power when sample sizes are all equal but that the LCL test will be more powerful in most cases where sample sizes are unequal across studies. The Stouffer test is expected to be more powerful only when sample and effect sizes have a substantial negative correlation.

A Simulation Study

A simulation study was conducted to verify, in some finite sample situations, the generalizations suggested by the large-sample theory.

Method

Design. The simulation study used a five-factor design systematically varying (a) the number of studies in the meta-analysis [$k = 12, 24,$ and 36], (b) the average value of the population effect size across studies [$\bar{\delta} = 0.0, 0.15,$ and 0.30], (c) the standard deviation of the population effect sizes across studies [$\sigma_\delta^2 = 0.0, 0.05,$ and 0.15], (d) the within-group sample sizes of the individual studies (see below), and (e) the correlation between the within-study sample sizes and the population effect sizes [$\rho = -0.3, 0.0,$ and 0.3]. The within-group sample sizes $n_i^E = n_i^C = n_i$ of the individual studies were assembled according to three patterns: (a) $n_1 = \cdots = n_6 = 20$, (b) $n_1 = n_2 = 10, n_3 = n_4 = 20, n_5 = n_6 = 30$, and (c) $n_1 = 5, n_2 = 10, n_3 = 15, n_4 = 25, n_5 = 30, n_6 = 35$. For $k = 12$, the pattern was repeated four times, and for $k = 36$, the pattern was repeated six times.

For sufficiently large values of sample size and effect sizes, the power of both tests is very close to the maximum possible value of one. In this range of sample and effect sizes, comparisons between tests would be misleading because it is impossible for either test to be substantially more powerful than the other. Thus comparisons between tests, if they are to be meaningful, must be made in situations in which tests have moderate power. These values of sample sizes, effect sizes, and numbers of studies were chosen to be realistic for meta-analyses, yet not so large as to yield power values that are essentially one for both tests.

With all factors crossed, the design has the possibility of $3^5 = 243$ cells or conditions. However, some of the conditions implied by combinations of factor levels are impossible. For example, when the variance of σ_δ^2 of effect sizes across studies is zero, the correlation of sample size and effect size must also be zero. Similarly, when the within-study sample sizes are all equal, this correlation must be zero. Thus only 153 conditions were possible. A minimum of 2,000 replications (2,000 meta-analyses) were generated with the parameters implied by each condition.

Data generation. For each replication within a condition, a set of 12, 24, or 36 t statistics were generated (depending on the level of k for that condition). Each t statistic corresponded to the results of a single "study." The value of t_i for a study with effect size δ_i and sample sizes n_i^E and n_i^C was generated as

$$t_i = \sqrt{\tilde{n}_i} X / \sqrt{Y/\nu}$$

where $X = \delta_1 + Z/\sqrt{\tilde{n}_i}$, $\nu = n_i^E + n_i^C - 2$, $Z \sim N(0, 1)$ and $Y \sim \chi^2_\nu$. The values of Z were generated using the International Mathematics Subroutine Libraries (IMSL) subroutine DRNNOR, and the values of Y were generated using the IMSL subroutine DRNCHI. The sample effect sizes were computed as

$$d_i = X/\sqrt{Y/\nu}.$$

The p values for the t statistics and the values of $z(p) = -\Phi^{-1}(p)$ needed to compute Z_S were computed using the IMSL subroutines DTDF and DNORIN, respectively.

Data analysis. The Stouffer test statistic Z_S and the LCL test statistic Z_{LCL} were computed for each replication. The data were then analyzed by means of the Statistical Analysis System (SAS) packaged programs to determine the number of statistics in each cell that exceeded the $\alpha = .05$ and .01 critical values. Because the pattern of results was similar at the two significance levels, we report results here only for $\alpha = .05$. The number of instances in which one statistic was significant but the other was not was also tabulated for each cell. To investigate the power of the LCL test in relation to that of the Stouffer test, we calculated the power curves for both tests at the .05 level of significance (one-tailed). Power differences were obtained by subtracting the observed proportion of times Z_S exceeded $Z_{.05} = 1.645$ from the observed proportion of times Z_{LCL} exceeded the same value. A 95% confidence interval was constructed around the difference value. In addition, we considered the proportion of times that the LCL test rejected the null hypothesis and the Stouffer test failed to reject the null hypothesis (denoted by $Z_{LCL>S}$), the proportion of times the Stouffer test rejected the null hypothesis and the LCL test failed to reject the null hypothesis (denoted by $Z_{S>LCL}$), and the proportion of the times both tests agreed to reject or not to reject the null hypothesis (denoted $Z_{S=LCL}$).

Comparisons of test procedures are complicated if the rejection rates of the tests are not the same (and ideally equal to the nominal significance level) when the null hypothesis is true. For example, a test that rejects more often than an exact test when the null hypothesis is true may also reject more often when the null hypothesis is false, but the comparison is not entirely fair. To ensure that empirical power comparisons presented below were fair in this sense, we investigated the rejection rates of the two tests under all of the conditions in our design for which the null hypothesis was true, that is, conditions in which every population effect size was zero. The proportion of the replications that led to rejection of the null hypothesis at the nominal $\alpha = .05$ level of significance did not differ from .05 for either the Stouffer or the LCL test under any of the nine conditions examined. These results are consistent with the fact that the Stouffer test is an exact test and that the LCL test, while not exact, has a rejection rate very close to the nominal under the null hypothesis. The latter conclusion is supported by rather extensive numerical investigations of the distribution of Z_{LCL} in small samples (Hedges, 1982; Hedges & Olkin, 1985).

Results

The results of the simulation study summarized in Table 1 confirm that the two tests generally have rather similar power although the LCL test is generally slightly more powerful. The two tests lead to the same decision (to reject or not to reject) in the majority of cases. When they do not lead to the same decision, the LCL test is usually much more likely to lead to the (correct) decision to reject the null hypothesis. Note that all of the differences between the empirical rejection rates of the two

Table 1
Summary of the Results of the Simulation Experiment Comparing the Power of the Lower Confidence Limit (LCL) and Stouffer Tests

Configuration	$\bar{\delta} = 0$	$\bar{\delta} = .15$	$\bar{\delta} = .30$
Balance across studies			
Equal sample sizes			
Power of the LCL test	.055	.719	.983
Power of the Stouffer test	.050	.704	.981
$Z_{LCL>S}$	.005	.014	.002
$Z_{S>LCL}$	.000	.000	.000
$Z_{S=LCL}$	.995	.986	.998
Three distinct sample sizes			
Power of the LCL test	.059	.718	.983
Power of the Stouffer test	.052	.690	.977
$Z_{LCL>S}$	.013	.042	.007
$Z_{S>LCL}$	.006	.013	.001
$Z_{S=LCL}$	.981	.945	.992
Six distinct sample sizes			
Power of the LCL test	.077	.699	.980
Power of the Stouffer test	.055	.654	.972
$Z_{LCL>S}$	.029	.065	.011
$Z_{S>LCL}$	.007	.021	.002
$Z_{S=LCL}$	.964	.914	.987
Correlation between sample size and effect size			
$\rho = -.3$			
Power of the LCL test	.022	.563*	.966
Power of the Stouffer test	.024	.567*	.962
$Z_{LCL>S}$	.004	.035	.008
$Z_{S>LCL}$	.006	.039	.005
$Z_{S=LCL}$	.990	.926	.987
$\rho = 0$			
Power of the LCL test	.056	.719	.983
Power of the Stouffer test	.050	.690	.978
$Z_{LCL>S}$	.010	.036	.006
$Z_{S>LCL}$	.004	.007	.001
$Z_{S=LCL}$	.986	.957	.993
$\rho = .3$			
Power of the LCL test	.145	.826	.993
Power of the Stouffer test	.093	.765	.984
$Z_{LCL>S}$	.058	.066	.009
$Z_{S>LCL}$	.042	.005	.000
$Z_{S=LCL}$	.900	.929	.991
Situations in which each test shows greatest superiority			
LCL test has greatest advantage[a]			
Power of the LCL test		.743	
Power of the Stouffer test		.606	
$Z_{LCL>S}$		.140	
$Z_{S>LCL}$		.003	
$Z_{S=LCL}$		.857	
Stouffer test has greatest advantages[b]			
Power of the LCL test		.550	
Power of the Stouffer test		.626	
$Z_{LCL>S}$		.013	
$Z_{S>LCL}$		.090	
$Z_{S=LCL}$		.897	

Note. All differences between tests, except those marked with an asterisk, are statistically significant at the .001 level of significance.
[a] This configuration is defined by $\bar{\delta} = .15$, $\sigma_\delta^2 = .15$, $\rho = .3$, $k = 12$, and the pattern of six distinct sample sizes. [b] This configuration is defined by $\bar{\delta} = .15$, $\sigma_\delta^2 = .15$, $\rho = -.3$, $k = 36$, and the pattern of six distinct sample sizes.

tests reported in the table are statistically significant ($p < .001$) unless noted by an asterisk.

We predicted that the LCL test would be more powerful than the Stouffer test when the sample sizes were equal. The first panel of Table 1 shows that when the samples sizes are the same, the LCL test has slightly greater power than the Stouffer test. When sample sizes are unequal, the LCL test has a slightly greater advantage in power over the Stouffer test. The maximum power difference, for the third pattern of sample sizes and $\bar{\delta} = .15$, was 4.5%. In this condition, $Z_{LCL>S} = 6.5\%$ and $Z_{S>LCL} = 2.1\%$.

We also predicted that the LCL test would be more powerful than the Stouffer test when the correlation between sample size and effect size was positive but that the Stouffer test could be more powerful than the LCL test when the correlation was negative. The second panel of Table 1 confirms this prediction. If the correlation between sample size and effect size is zero or positive, the LCL test is slightly more powerful than the Stouffer test. When the correlation is negative, however, the power of the LCL test exceeds that of the Stouffer test when $\bar{\delta} = .30$. Thus, even when analytic considerations suggest that the Stouffer test should have the greatest advantage over the LCL test, it is significantly more powerful for small average effect sizes, significantly less powerful for large average effect sizes, and more powerful, but not significantly so, for intermediate average effect sizes. Moreover, the magnitude of the power advantage of the Stouffer test when $\rho = -.3$ is smaller than that of the LCL test when $\rho = 0$ or $\rho = .3$. For example, when $\rho = -.3$ and $\bar{\delta} = .15$, the power of the Stouffer test exceeds that of the LCL test by 0.4%, $Z_{S>LCL} = 3.9\%$, and $Z_{LCL>S} = 3.5\%$. But when $\rho = .3$ and $\bar{\delta} = .15$, the power of the LCL test exceeds that of the Stouffer test by 6.1%, $Z_{LCL>S} = 6.6\%$, and $Z_{S>LCL} = 0.5\%$.

Finally, we examined all 153 conditions to discover when the maximum differences between the two tests occurred. The results for the condition in which the LCL test exhibited the greatest superiority are reported in the third panel of Table I. When $\bar{\delta} = .15$, $\sigma_\delta^2 = .15$, the sample sizes show six distinctions (the third pattern), $\rho = .3$, and $k = 12$, the power of the LCL test exceeded that of the Stouffer test by 13.7%, $Z_{LCL>S} = 14.0\%$, and $Z_{S>LCL} = 0.3\%$. The Stouffer test exhibited the greatest superiority when $\bar{\delta} = .15$, $\sigma_\delta^2 = .15$, the sample sizes show six distinctions, $\rho = -.3$, and $k = 36$. In this condition, the power of the Stouffer test exceeded that of the LCL test by 7.6%, $Z_{S>LCL} = 9.0\%$, and $Z_{LCL>S} = 1.3\%$.

Conclusion

The Stouffer test of the significance of combined results and the LCL test of the significance of the weighted average effect size can both be viewed as tests of essentially the same null hypothesis. That is, both can be viewed as tests of the null hypothesis that the effect size is zero in every study or that the average effect size is zero. Our results suggest, therefore, that there is no justification for computing both the Stouffer test and a test that the mean effect size differs from zero.

One rather formal criterion for choosing between the two tests is statistical power. The power of the two tests does not differ substantially in many situations. When they do differ in

power, both analytic results and our simulation results suggest that the LCL test is usually slightly more powerful than the Stouffer test. The most substantial advantages in power for the LCL test occur where effect sizes are modest and study sample sizes vary substantially, so that studies with larger sample sizes have larger effect sizes. However, when sample sizes are negatively correlated with effect sizes (small studies have larger effects), the Stouffer test may be slightly more powerful than the LCL test. Of course neither test will be superior when sample sizes or effect sizes are very large. In such cases, both tests will reject the null hypothesis nearly 100% of the time. Consequently, power considerations suggest that the LCL is generally preferable whenever it can be applied, that is, whenever effect size estimates are available from each study.

Another criterion for choosing a test procedure is the clarity of the relationship between the test and a meaningful estimate of effect magnitude. This criterion is less formal than broadly conceptual. Its importance depends on the importance attached to estimation in the interpretation of research results. The role of effect magnitude is hidden in the mathematics of the Stouffer procedure. The average effect magnitude is explicit in the LCL test.

A final consideration has less to do with comparing tests than choosing statistical analysis strategy. Despite calls for the increased use of estimation, hypothesis-testing strategies still seem to predominate in primary research. Estimation of effects has usually, but not always, predominated in meta-analysis (Glass et al., 1981). Part of the reason may be that strategies involving estimation are frequently more informative than null-hypothesis-testing strategies (see Becker, 1987; Hedges & Olkin, 1985). Given that estimation strategies generally have advantages of greater interpretability, the LCL test has another advantage. The weighted mean effect size and its standard error are (or can be) computed from the same components required to compute Z_{LCL}. Thus an estimation of the average effect size, its standard error, and a confidence interval for effect size may be computed with very little additional effort.

References

Becker, B. J. (1985). *Applying tests of combined significance: Hypotheses and power considerations.* Unpublished doctoral disseration, University of Chicago.

Becker, B. J. (1987). Applying tests of combined significance. *Psychological Bulletin, 102,* 164–171.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society 4*(Suppl. 4), 102–118.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). San Diego, CA: Academic Press.

Cooper, H. M. (1989). *Integrative research: A guide for literature reviews.* Newbury Park, CA: Sage.

Fisher, A. (1932). *Statistical methods for research workers* (4th ed.). London: Oliver & Boyd.

Glass, G. V, McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Hedges, L. V. (1982). Estimating effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Lambert, D. M. (1978). *P-values: Asymptotics and robustness.* Unpublished doctoral disseration, University of Rochester, Rochester, NY.

Lambert, D. M., & Hall, W. J. (1982). Asymptotic lognormality of P-values. *Annals of Statistics, 10,* 44–64.

Lambert, D. M., & Hall, W. J. (1983). Correction to "Asymptotic lognormality of *P*-values." *Annals of Statistics, 11,* 348.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185–193.

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during army life* (Vol. 1). Princeton, NJ: Princeton University Press.

Received August 17, 1990
Revision received May 8, 1991
Accepted June 14, 1991 ■