Psychological Methods
2001, Vol. 6, No. 4, 371-386

1082-989X/01/$5.00 DOI: 10.1037//1082-989X.6.4.371

Evaluating Statistical Difference, Equivalence, and Indeterminacy Using Inferential Confidence Intervals: An Integrated Alternative Method of Conducting Null Hypothesis Statistical Tests

Warren W. Tryon
Fordham University

Null hypothesis statistical testing (NHST) has been debated extensively but always successfully defended. The technical merits of NHST are not disputed in this article. The widespread misuse of NHST has created a human factors problem that this article intends to ameliorate. This article describes an integrated, alternative inferential confidence interval approach to testing for statistical difference, equivalence, and indeterminacy that is algebraically equivalent to standard NHST procedures and therefore exacts the same evidential standard. The combined numeric and graphic tests of statistical difference, equivalence, and indeterminacy are designed to avoid common interpretive problems associated with NHST procedures. Multiple comparisons, power, sample size, test reliability, effect size, and cause–effect ratio are discussed. A section on the proper interpretation of confidence intervals is followed by a decision rule summary and caveats.

The long-standing controversy surrounding null hypothesis statistical testing (NHST) has typically been argued on its technical merits, and they are not discussed here. I focus on the serious human factors problem that the well-documented widespread misuses of NHST procedures have created for research producers and consumers, and I recommend new inferential confidence interval (CI) procedures that are designed to minimize these misuses. The first section of this article briefly reviews the human factors problem at hand and proposes reasons for it that form the basis for subsequent sections that introduce procedures for making decisions about statistical difference, equivalence, and indeterminacy using both graphic and numeric methods. Redundancy between graphic and numeric methods combined with clear decision criteria should facilitate an arrival at correct conclusions more consistently than has occurred with standard NHST procedures. A section on what CIs mean will hopefully limit their misinterpretation. I end with a decision rule summary and caveats.

The Human Factors Problem

Sir Ronald Fisher[1] introduced NHST in 1925, and it has since become the cornerstone of statistical analysis in social science research. Pearce (1992) reported that criticism of NHST procedures began immediately on Fisher's introduction of them. Criticism of NHST has continued unabated for the past 75 years

I thank Larry DeCarlo for bringing the Goldstein and Healy (1995) article to my attention, for reviewing my statistical assertions, for making specific textual comments, and for our general discussions about statistical difference. I thank Herman Friedman for bringing the Pearce (1992) article to my attention. I thank Tenko Raykov for reviewing a draft of this article. I thank Charles Lewis for confirming the confidence interval interpretations made in this article. Finally, I thank Anne Anastasi for encouraging my efforts to develop confidence interval alternatives to null hypothesis statistical test procedures.

Correspondence concerning this article should be addressed to Warren W. Tryon, Department of Psychology, Fordham University, 441 East Fordham Road, Bronx, New York 10458-5198. Electronic mail may be sent to wtryon@fordham.edu.

[1] Arbuthnot (1710) used statistics to test the hypothesis that more boys than girls are born. He observed that more boys than girls were born in each of 82 consecutive years for which he had records. The probability of this event occurring by chance is a miniscule $(1/2)^{82} = 2.0 \times 10^{-25}$. (See Glass & Hopkins, 1970, pp. 255–256).

(cf. Bakan, 1966; Berkson, 1942; Carver, 1978, 1993; Chow, 1998; Cohen, 1994; Dar, Serlin, & Omer, 1994; Hagen, 1997; Harlow, 1997; Hodges & Lehmann, 1954; Hogben, 1957; Hunter & Schmidt, 1990; Lykken, 1968; Meehl, 1967, 1978, 1990a, 1990b; Morrison & Henkel, 1970; Rozeboom, 1960; Sterling, 1959) and culminated in a special section of *Psychological Science* discussing whether the NHST should be banned (Abelson, 1997b; Estes, 1997b; Harris, 1997b; Hunter, 1997; Scarr, 1997; Shrout, 1997). The American Psychological Association Task Force on Statistical Inference was convened to determine what role, if any, NHST should have in psychological science (Schmidt, 1996; Wilkinson and the Task Force on Statistical Inference, 1999). Harlow, Mulaik, and Steiger (1997) presented a book-length debate about significance tests and alternatives. Nickerson (2000) provided a comprehensive critical review of NHST issues.

The technical merits of the NHST are not disputed here. NHST proponents have successfully defended NHST procedures when correctly used (Abelson, 1997a, 1997b, Hagen, 1997; Harris, 1997a, 1997b; Mulaik, Raju, & Harshman, 1997; Rindskopf, 1997; Serlin, 1993; Serlin & Lapsley, 1985). Krantz (1999) observed that "statisticians prove theorems or develop methods that, if properly applied, would be useful. If people misapply them, this is viewed as a problem for education, not for statistical research" (p. 1374).

The problem identified here is that NHST procedures have long been and currently are frequently misused, as Krantz (1999), Nickerson (2000), and others have acknowledged. Dar et al. (1994) reported that prominent investigators publishing in our best peer-reviewed journals for at least 3 decades, as of 1994, have consistently misused NHST procedures. It follows that the journal editors and reviewers who published these articles did not catch these mistakes,[2] perhaps because they similarly misuse NHST procedures. One or more aspects of NHST procedures are mistaught in at least six books written by leading psychometricians (Cohen, 1994). Hagen's (1997) need to revise three of Cohen's (1994) NHST criticisms indicates that even a prominent author of advanced statistics textbooks cannot always correctly interpret NHST results. Authors of nearly two dozen introductory psychology texts published between 1965 and 1994 err in their presentation of NHST procedures (McMan, 1995). Nickerson (2000) examined what he considered to be the more important and fundamental misconceptions and false beliefs concerning NHST procedures. Schmidt and Hunter (1997) described eight common misunderstandings about NHST procedures that are used as reasons why NHST practices should not continue. If so many well-trained and distinguished investigators publishing in leading psychology journals, the editors and reviewers of these journals, and the authors of psychometrics textbooks cannot consistently use NHST procedures correctly, then it is reasonable to expect that less well-trained academicians, investigators, and graduate students may make similar mistakes when using standard NHST procedures. Such widespread misuse of a fundamental procedure for evaluating data cannot be constructive. We simply must find a way to ameliorate this human factors problem.

Substantive and repeated efforts made over the past 75 years to remediate NHST misuses by discussing the technical merits and correct use of NHST have not been productive (e.g., Carver, 1978, 1993; Kaiser, 1960; Kish, 1959; Serlin & Lapsley, 1985). There is no reason to expect that further efforts along these lines will be useful. A new approach to implementing NHST procedures designed to avoid interpretive errors of the past seems worth trying, given the importance of correct statistical inference and the ineffectiveness of prior educational efforts. The approach developed here is based on the premise that the primary difficulty with standard NHST stems from misunderstandings about the null hypothesis. NHST procedures calculate the probability of data given that the null hypothesis is true, $p(D|H_0)$, yet investigators want to know the probability that the null hypothesis is true given the data, $p(H_0|D)$. Conclusions about the latter erroneously follow from calculations of the former.[3] Cohen (1994) called this NHST misuse "the permanent illusion" (p. 998). The null hypothesis is perceived as the hand of chance. Rejecting it therefore seems to reject all chance interpretations. The allied replication misunderstanding is that results are repli-

[2] Raymond S. Nickerson (personal communication) correctly observed in his review of this article that there is not complete agreement among experts in the field regarding what constitutes a mistake in all instances. Even those who developed NHST procedures (e.g., Fisher, Neyman, Pearson) did not agree, and their differences have never been reconciled.

[3] Markus (2001) discussed this converse inequality problem in detail. Nickerson (2000, p. 252) demonstrated that, under some conditions, $p(H_0|D)$ can be similar to $p(D|H_0)$.

cable if they are not due to chance. Hence, rejecting the null hypothesis appears to certify replicability. This ignores the possibility that nonrandom differences can arise from artifact, legitimate effects due to unknown causes as well as from hypothesized reasons. The fact that a hypothesis passes a statistical test does not ensure that someone else can replicate methods, let alone results.

Two additional common NHST misuses stem from nonsignificant findings. It is common practice to interpret "marginally" significant results as trends and then discuss them as real differences. When the data are in the predicted direction and the p value is near the .05 level, accepting the null hypothesis seems overly harsh. The results seem more in line with difference than with no difference, so investigators emphasize the former conclusion over the latter. Access to a third alternative provides a more palatable solution; namely, that the results are statistically indeterminate. This alternative has always been technically available, but standard NHST procedures emphasize accepting or rejecting the null hypothesis. Harris (1997a) and Kaiser (1960) identified the need for a third alternative. The methods proposed in this article clearly incorporate statistical indeterminacy and should make this alternative more salient.

The other use to which nonsignificant results are erroneously put is to conclude that two groups are matched or even equivalent when an NHST fails to find a difference. The same mistake is made in comparisons of statistics. For example, a finding that test reliability is not significantly different across groups of subjects or across test administrations to the same subjects is erroneously interpreted as equivalent test reliability across these dimensions. The section on statistical equivalence in this article explains why this is not so and provides a method for properly reaching such conclusions.

A final misuse of NHST procedures pertains to reporting p values. Serlin (1993) noted that to correctly conduct NHST and properly establish grounds for falsification, one must specify the Type I alpha level (e.g., .05, .01) in advance of conducting the analysis. This classic NHST step is rarely, if ever, taken; its omission may constitute the most common NHST misuse. Investigators typically select the p value after computing the test statistic and then report the highest significance level, the lowest p value that their inferential test statistic allows. For example, $t(30) = 3.70$ would be reported as $p < .001$ because it exceeds the critical values for the .05, .01, and .001 significance levels. All of the major software packages report significance levels in this way. None of them require the user to establish a level of statistical significance as a necessary condition for calculating test statistics and then report results as significant or not. The recommended procedures avoid this error.

In sum, the well-documented misuses of NHST procedures largely entail the null hypothesis. Although there is no way to avoid the null hypothesis when conducting NHST, the procedures recommended in this article seek to minimize NHST misuses by shifting attention away from the null hypothesis toward inferences about statistical difference, equivalence, and indeterminacy between or among group means. A corollary premise is that presenting statistical difference, equivalence, and indeterminacy in a single integrated data analytic context will help curb NHST misuses. The use of redundant graphic and numeric methods is also expected to be helpful. The recommended computational procedures will largely eliminate the p-value reporting misuses I have noted.

Statistical Difference

This section introduces the concept of *inferential CI*. The first subsection applies this concept to two independent means, and the next to two dependent means, with numerical examples. Statistical difference[4] is reported when two inferential CIs do not

[4] Scarr (1997) correctly noted that the term *statistical significance* is frequently confused with its ordinary English usage, meaning *importance,* and recommended replacing it immediately. Her preferred replacement term, *reliability,* is problematic because it suggests that $p < .05$ implies replicability of $p > .95$ on the premise that rejection of the null hypothesis implies that the results are not due to chance and that therefore they must be both systematic and reproducible. Carver (1978) identified these views as common misunderstandings of NHST. Significance testing does not provide evidence of reproducibility. Carver (1978) indicated that reproducibility depends on the extent to which other investigators can manipulate and control relevant variables. Passing an NHST is not evidence that other investigators will succeed in duplicating the reported outcome. Carver (1978) was especially concerned that this misunderstanding of NHST would reduce efforts to replicate results and that failures to replicate would not be published because they do not pass the NHST. *Statistical difference* is a preferable term to *statistical significance* because it is descriptive and without the undesired connotation of importance. It is note-

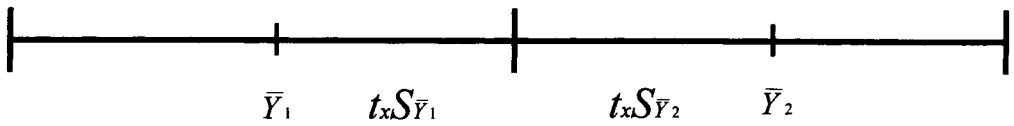

Figure 1. Confidence interval test model: Two nonoverlapping but abutted confidence intervals.

overlap. Tests and conclusions about statistical equivalency and indeterminacy follow in subsequent sections.

Several authors have suggested that decisions about statistical difference can be reached using *descriptive* CIs and that their use is less subject to abuse than are standard NHST procedures (Abelson, 1997b; Andrews, Snee, & Sarner, 1980; Cohen, 1994; Estes, 1997a; Gabriel, 1978; Hochberg, Weiss, & Hart, 1982; Hunter, 1997; Loftus, 1991, 1993, 1995, 1996; Loftus & Mason, 1994; McGill, Tukey, & Larsen, 1978; Rosenthal, 1995; Serlin, 1993; Tukey, 1991). Establishing a descriptive 95% CI about the differences between matched pairs implements the same evidential standard as a t test does. However, establishing descriptive 95% CIs about each of two means and concluding that the means differ, $p < .05$, if the two CIs do not overlap constitutes a substantially greater burden of proof of statistical difference than does the t test. Below, I recommend an inferential CI approach that corrects this problem.

Goldstein and Healy (1995) demonstrated that 95% CIs can overlap substantially and yet the two means will be significantly different at the 5% level. They demonstrated that if the standard errors of two independent means are equal, then nonoverlapping 84% CIs are equivalent to a Z test of these means at the .05 level. More precisely, adding and subtracting 1.39 standard errors[5] to and from each mean establishes a CI that can be used to certify $p < .05$ differences if the CIs do not overlap. This amounts to a reduction of the width of the standard 95% CI about each mean to 71% (1.39/1.96) of its normal size. The actual reduction depends on the sample size and standard deviation for each group. Put otherwise, the nonoverlapping descriptive 95% CI method is at least 1.96/1.39 = 1.41 or 41% more stringent than the standard Z test and therefore is not a suitable NHST alternative. Hence, direct comparison of descriptive 95% CIs constitutes a more rigorous standard of statistical difference than an ordinary t test. An important benefit of the Goldstein and Healy (1995) method is that it provides a basis for the comprehensive inferential CI alternative NHST advocated below.

Two Independent Means

The recommended procedure derives from Goldstein and Healy (1995), who described a graphical method that enables one to simultaneously discern which independent means are statistically different from other means by constructing a modified CI, here called an inferential CI, about each mean. The procedure reduces the standard descriptive CI such that nonoverlapping inferential CIs are algebraically equivalent to an NHST. It also provides a context in which one can test for statistical equivalence and indeterminacy. Goldstein and Healy (1995) described their method in the context of a large-sample Z test. I now extend their work to the t test. The CI for the mean is given in Equation 1 as the mean plus and minus t standard errors ($df = n - 1$) of the mean, where the standard error of the mean equals the standard deviation divided by the square root of the sample size:

$$\bar{Y} \pm t_{\alpha/2} S_{\bar{Y}} = \bar{Y} \pm t_{\alpha/2} \frac{S}{\sqrt{N}}. \qquad (1)$$

Figure 1 illustrates two nonoverlapping but abutted CIs about two means. The difference between two means equals the sum of the half-widths of the two CIs. Our objective is to determine the t value (t_x) that makes the two means significantly different at the stated significance level.

Equation 2 begins with the equation for the independent t test, substitutes the CI expression for the difference between the two means depicted in Figure

worthy that a misinterpretation of NHST identified in 1978 is recommended as a solution to an NHST nomenclature problem in 1997.

[5] The area under the normal curve between the mean and $Z = 1.39$ is .4177. Doubling this area gives .8354, or 84% confidence.

1, factors out the t value needed to ensure that two nonoverlapping CIs will be statistically significant at the stated probability level (t_x), and reexpresses the standard error of the difference between two means in terms of the standard error of each group:

$$t_{95} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{\bar{Y}_1 - \bar{Y}_2}} = \frac{t_x S_{\bar{Y}_1} + t_x S_{\bar{Y}_2}}{S_{\bar{Y}_1 - \bar{Y}_2}} = \frac{t_x (S_{\bar{Y}_1} + S_{\bar{Y}_2})}{S_{\bar{Y}_1 - \bar{Y}_2}} = t_x \frac{S_{\bar{Y}_1} + S_{\bar{Y}_2}}{\sqrt{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_2}}}. \quad (2)$$

This equation holds for both equal and unequal sample sizes because it depends on individually computed standard errors equal to the sample standard deviation divided by the square root of its sample size.

Equation 3 has been solved for t_x and defines E as the ratio of the standard error of the difference between two groups to the sum of the standard errors of both groups:

$$t_x = t_{95} \frac{\sqrt{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_2}}}{S_{\bar{Y}_1} + S_{\bar{Y}_2}} = t_{95} E. \quad (3)$$

This ratio can be visualized for two independent groups as the length of the hypotenuse of a right triangle divided by the sum of the lengths of the two legs. The length of each leg corresponds to the standard error of the mean for each group. The length of the hypotenuse corresponds to the standard error of the difference between the two means. This ratio also holds for two dependent groups where the angle between the two legs is less than 90°, as specified by the law of cosines and the geometric equivalence between correlation and cosines (cf. Guilford & Fruchter, 1973, p. 151; Nunnally, 1967, pp. 299–300). The symbol E (for experimental design) is introduced here to draw attention to the fact that the tabled value of the t distribution is being reduced by a composite factor that is a function of the experimental design. E equals a minimum, in an independent-groups design, of $\sqrt{2}/2$ when the standard errors are equal, making t_x = .7071 t_{95}. E gradually increases toward unity as the standard errors become increasingly unequal, either because of unequal within-group variances or because of unequal sample sizes. However, even when one standard error is as much as five times larger than the other, E only increases to .8498. Good experimental design favors the minimum possible value for E and, therefore, equal standard errors. However, the cost of recruiting subjects in different groups may not be the same. The behavior of E is independent of significance level (i.e., t value). Equation 4 summarizes the construction of a reduced inferential CI about each mean so that nonoverlap equates to statistical difference by a standard t test:

$$\bar{Y} \pm E\, t_{\alpha/2}\, S_{\bar{Y}} = \bar{Y} \pm E\, t_{\alpha/2} \frac{S}{\sqrt{N}}. \quad (4)$$

These results can be expressed numerically and/or graphically.

Table 1 contains simulated data that I use to illustrate the recommended procedures and some of their properties. The raw data are in the Appendix. Suppose that Groups A, B, and C are independent of each other. We now test Groups A and B for statistical difference. The first step is to establish the level of statistical significance we should work at to determine how to calculate descriptive CIs about each mean. We choose the 5% level of statistical significance for 95% confidence. Next we calculate E, the extent to which the descriptive CI must be reduced to obtain an inferential CI on the basis of the experimental design used. Substituting the standard errors into part of Equation 3 yields

$$E = \frac{\sqrt{(2.6920)^2 + (0.7054)^2}}{2.6920 + 0.7054} = .8191.$$

The inferential CI is 81.91% as large as the descriptive CI. The critical t value for $df = N - 1 = 19$ at the 5% significance (95% confidence) level in this case is 2.0930. The inferential CI is constructed using a proportionately reduced critical t value. The reduced t value is t_x = .8191 (2.0930) = 1.7144. The resulting inferential CIs for each of the two groups are as follows:

Group A: $68 \pm 1.7144\ (2.6920) = 63.385$ to 72.615

Group B: $75 \pm 1.7144\ (0.7054) = 73.791$ to 76.209.

Statistical difference is said to exist between the two groups because the two inferential CIs do not overlap; the upper limit of the lesser mean (72.615) is less than the lower limit of the greater mean (73.791). The probability value associated with this statistical difference is $p < .05$ because the critical value for the 5% significance level (95% confidence level) was the initial t value that was reduced by E to obtain the inferential CI. We would have had to have chosen the t value for constructing a 99% CI to test for difference at the .01 significance level. Notice that this procedure

Table 1
Descriptive Statistics for Simulated Data for Four Independent Groups

Statistic	Group A	Group B	Group C	Group D
M	68.0	75.0	76.0	71.0
SD	12.0	3.2	2.4	11.0
N	20	20	20	25
df	19	19	19	24
t(*df*)	2.0930	2.0930	2.0930	2.0639
SE	2.6920	0.7054	0.5293	2.1921
Lower 95% CI	68.0 − 2.093 (2.6920) = 62.366	75.0 − 2.093 (0.7054) = 73.523	76.0 − 2.093 (0.5293) = 74.892	71.0 − 2.064 (2.1921) = 66.476
Upper 95% CI	68.0 + 2.093 (2.6920) = 73.634	75.0 + 2.093 (0.7054) = 76.476	76.0 + 2.093 (0.5293) = 77.108	71.0 + 2.064 (2.1921) = 75.524

Note. CI = confidence interval.

requires the user to establish the significance level before calculating the test statistic, the inferential CIs.

Suppose that one planned to compare Group A with Groups B and C, thereby conducting two tests of statistical difference. One way to control the experimentwise Type I alpha error rate is to use the t value associated with two comparisons. Other methods of controlling Type I error associated with multiple comparisons are discussed below. Because all means are based on 20 subjects, each mean is associated with 19 degrees of freedom. A Bonferroni corrected table, such as Table D.15 in Winer, Brown, and Michels (1991, pp. 992–995), gives the critical t value for 19 degrees of freedom for two comparisons at the 5% level as 2.4334. This makes t_x = .8191 (2.4334) = 1.9932, resulting in slightly wider inferential CIs for the two groups. Evidence of statistical difference remains, as these two inferential CIs do not overlap.

Group A: $68 \pm 1.9932\ (2.6920) = 62.634$ to 73.366

Group B: $75 \pm 1.9932\ (0.7054) = 73.594$ to 76.406.

The need to use larger t values for additional comparisons directly reminds the user to judiciously choose the relevant tests of statistical difference. At some point, the required t value may become large enough that evidence of statistical difference is no longer present. This would have occurred in this example had three comparisons been chosen; the critical t value for three comparisons on 19 degrees of freedom is 2.6251. The recommended procedures discourage the routine use of making all possible pairwise comparisons among means, and they motivate investigators to make as few comparisons as their research allows.

A second example illustrates calculations when sample sizes are unequal. We compare the means of Groups A and D in Table 1. We calculate the standard error for each group by dividing the standard deviation for each group by the square root of the number of subjects in that group, thereby accommodating unequal sample size across groups. E is calculated as follows:

$$E = \frac{\sqrt{(2.6920)^2 + (2.1921)^2}}{2.6920 + 2.1921} = .7108.$$

A separate value for t_x is calculated for each group because the critical t value for the 5% significance level depends on sample size, which is different in the two groups. Group A contains 20 subjects and therefore 19 degrees of freedom for which the critical t

value is 2.0930. Group D contains 25 subjects and therefore 24 degrees of freedom for which the critical t value is 2.0639. This yields the following two values of t_x:

Group A: $t_x = .7108\ (2.0930) = 1.4877$

Group D: $t_x = .7108\ (2.0639) = 1.4670.$

The corresponding inferential CIs are

Group A: $68 \pm 1.4877\ (2.6920) = 63.995$ to 72.005

Group D: $71 \pm 1.4670\ (2.1921) = 67.784$ to $74.216.$

Notice that the inferential CIs overlap. These two means are therefore not statistically different. Neither do we yet have evidence of statistical equivalence (see below). Consequently, we conclude that statistical indeterminacy exists at this point in the analysis. Statistical indeterminacy precludes any conclusion about these two means. Such uncertainty is the indeterminacy at issue. We cannot infer that the means are equivalent or matched until we conduct a test for statistical equivalence and obtain positive results. Absence of positive evidence of difference is not positive evidence of equivalence.

Multiple comparisons require the use of a larger critical t value. Correspondingly, larger t values are selected for each group. For example, the critical t value for making two comparisons with $N = 20$, $df = 19$ is 2.4334 versus the critical t value for making one comparison of 2.0930 used above. Similarly, the critical t value for making two comparisons with $N = 25$, $df = 24$ is 2.3909, as opposed to the critical t value for making one comparison, 2.0639. The same E value of .7108 is used because the standard errors have not changed. The new t_x values become

Group A: $t_x = .7108\ (2.4334) = 1.7297$

Group D: $t_x = .7108\ (2.3909) = 1.6995.$

The corresponding inferential CIs are

Group A: $68 \pm 1.7297\ (2.6920) = 63.344$ to 72.656

Group D: $71 \pm 1.6995\ (2.1921) = 67.275$ to $74.725.$

These overlapping inferential CIs support an inference of statistical indeterminacy. Ordinarily, there would be no point to computing inferential CIs for two or more comparisons when statistical indeterminacy was found for a single comparison, because the larger critical t values would only confirm statistical indeterminacy. I make these calculations here only to illustrate how they are accomplished when sample sizes are unequal.

Two Dependent Means

The previous analysis for two independent means can be extended to two dependent means of equal or unequal sample sizes. The E term has the same meaning in Equation 5 as it does in Equation 4. The only difference between Equations 4 and 5 is the presence of a term that is twice the correlation between the two variables multiplied by the standard errors of each variable. The correlation between the two variables further reduces E below what is possible in the independent-groups design, thereby further narrowing the inferential CIs about each mean:

$$t_x = t_{95} \frac{\sqrt{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_2} - 2r_{12}\, S_{\bar{Y}_1}\, S_{\bar{Y}_2}}}{S_{\bar{Y}_1} + S_{\bar{Y}_2}} = E\, t_{95}. \quad (5)$$

Table 2 contains simulated data for two dependent groups (Groups E and F). Notice that the descriptive 95% CIs overlap. The correlation between these two data sets is .521, and I use this as the estimate of the population correlation. The E coefficient shows that the critical t value must be reduced to .4899, less

Table 2
Descriptive Statistics for Simulated Data for Two Dependent Groups

Statistic	Group E	Group F
M	50.0	54.0
SD	5.3	5.6
N	20	20
df	19	19
t(*df*)	2.0930	2.0930
SE	1.1916	1.2552
Lower 95% CI	50.0 − 2.093 (1.1916) = 47.506	54.0 − 2.093 (1.2552) = 51.373
Upper 95% CI	50.0 + 2.093 (1.1916) = 52.494	54.0 + 2.093 (1.2552) = 56.627

Note. CI = confidence interval.

than half its original value, for us to establish inferential CIs:

$$E = \frac{\sqrt{(1.1916)^2 + (1.2552)^2 - 2\,(0.521)(1.1916)(1.2552)}}{1.1916 + 1.2552} = .4899.$$

The critical t value for $N = 20$, $df = 19$, at the 5% significance level is 2.0930. This results in a value of $t_x = .4899\ (2.0930) = 1.0254$ and the following two inferential CIs:

Group E: $50 \pm 1.0254\ (1.1916) = 48.778$ to 51.222

Group F: $54 \pm 1.0254\ (1.2552) = 52.713$ to 55.287.

Evidence of statistical difference exists because these two inferential CIs do not overlap. Again, the p value associated with this statistical difference is .05 because we selected the critical t value for the 5% significance level before multiplying by E. The additional reduction in E derived by using the correlation between the two groups is authorized by the use of a dependent-groups experimental design. One is not entitled to calculate and use such correlations when an independent-groups experimental design has been selected.

Multiple Comparisons

The problem of multiple comparisons is addressed here in two ways. The preferred approach is to use the standard Bonferroni method for limiting experimentwide alpha error when conducting multiple tests (cf. Winer et al., 1991, pp. 158–165). One selects a t value depending on the number of tests one plans to make. More comparisons require a larger t value. This value is reduced when it is multiplied by the E factor to get t_x, the number of standard errors that we add to and subtract from the mean to establish the inferential CI about that mean.

Another approach is to calculate an average E over all possible pairwise combinations of standard errors involved in the planned comparisons. This procedure has the advantage of creating a single set of intervals that can be compared with each other either quantitatively or graphically (cf. Goldstein & Healy, 1995). The disadvantage is that the average E will be larger than some individual Es, resulting in correspondingly less powerful tests for those comparisons. On the other hand, comparisons whose individual E is larger than the average E are more powerful with this approach.

Keselman, Cribbie, and Holland (1999) reviewed and compared four prominent methods of controlling Type I error stemming from multiple comparisons (cf. Williams, Jones, & Tukey, 1999). Methods recommended by Benjamini and Hochberg (1995), Hochberg (1988), and Shaffer (1986, 1995) begin by ordering comparisons on the basis of their p values. It is noteworthy that these methods depend on the NHST p value misuse discussed previously, the practice of reporting the highest level of statistical significance that the test statistic can support after computing the test statistic rather than selecting a significance level before computing the test statistic and then either rejecting or accepting the null hypothesis. Opposing this NHST abuse precludes using these procedures. On the other hand, the benefits of these procedures may suggest deviation from classic NHST practices. The commitment here to implement an exact NHST parallel using inferential CIs favors the first method described.

Statistical Equivalence

A common misuse of NHST is to interpret the absence of statistical significance as the presence of statistical equivalence despite a substantial literature to the contrary (e.g., Anderson & Hauck, 1983; Blackwelder, 1982; Detsky & Sackett, 1985; Dunnett & Gent, 1977; Makuch & Simon, 1978; Metzler, 1979; Rogers, Howard, & Vessey, 1993; Selwyn, Dempster, & Hall, 1981; Selwyn & Hall, 1984; Westlake, 1972, 1976, 1979, 1981). Frick (1995) sought to establish statistical equivalence by accepting the null hypothesis under certain conditions, including the subjective criterion of having made a good effort to find an effect. Findings of no difference can be theoretically and practically important, but good evidence for such claims is better provided by methods that establish statistical equivalence than by good failed efforts at finding statistical difference. The flawed logic appears to be (a) either two means (proportions, correlations, variances, etc.) are different or they are not, (b) they are not significantly different, (c) hence, they are equivalent. Alternatively, (a) the null hypothesis assumes that the two means (proportions, correlations, variances, etc.) are identical, (b) the null hypothesis has not been rejected, (c) hence, the two means are statistically equivalent. Conclusions about statistical equivalence are more frequently implicit than ex-

plicit.[6] For example, two groups are considered matched if they are not significantly different. Reliability coefficients in two groups are considered to be the same if they are not statistically different. Theories that predict no difference are viewed as supported if the differences in question are not statistically significant. The mistake in all these cases is to accept the null hypothesis as true when insufficient evidence exists to reject it. Blackwelder (1982) correctly noted that

> it is inappropriate to base a conclusion that therapies are equivalent on whether the observed significance level (*p* value) for the null hypothesis of equality is larger than some arbitrary small value, such as 0.05 or 0.01; *p* is a measure of the evidence against the null hypothesis, not for it, and insufficient evidence to reject the null hypothesis does not imply sufficient evidence to accept it. (p. 346)

Said otherwise, absence of positive evidence for statistical difference does not constitute presence of positive evidence for statistical equivalence. This reasoning extends to the procedures described above in that the overlap of two inferential CIs is not sufficient evidence of statistical equivalence, although it is a necessary condition for statistical equivalence. I now extend the procedures described above to correctly inferring statistical equivalence.

Dunnett and Gent (1977) established statistical equivalence by calculating the descriptive CI for the difference in question (lowercase delta, δ) and determining whether it lies within an amount that is considered inconsequential (uppercase Delta, Δ) on the basis of substantive theoretical considerations. Said otherwise, the 95% descriptive CI for an observed difference, δ, must lie entirely within a band of no consequence, ± Δ / 2. In other words, the entire 95% CI for delta (δ) must lie below the high end of the Delta (Δ) interval. The lower limit of the 95% CI for delta (δ) frequently lies below the high end of the Delta (Δ) interval and therefore is rarely an issue. The main consideration is whether the upper limit of the 95% CI for delta (δ) also lies below the high end of the Delta (Δ) interval.

Multiplying both sides of Equation 3 by $S_{\bar{Y}_1} + S_{\bar{Y}_2}$ and adding $d = \bar{Y}_1 - \bar{Y}_2$ to both sides gives Equation 6:

$$d + t_x(S_{\bar{Y}_1} + S_{\bar{Y}_2}) = d + t_{95}\sqrt{S^2_{\bar{Y}_1} + S^2_{\bar{Y}_2}}. \quad (6)$$

The right side of Equation 6 is the upper limit of the standard 95% CI of the observed effect. Hence, when it is below the high end of the Delta (Δ) interval, the left side of the equation is also below the high end of the Delta interval. The left side of Equation 6 can be reformulated as $t_x S_{\bar{Y}_1} + d + t_x S_{\bar{Y}_2}$, which corresponds to what is defined in Figure 2 as the range (R_g) of two overlapping, appropriately reduced inferential CIs. R_g begins with the lower CI limit of the lesser mean, ($\bar{Y}_1$) in this case, and extends to the upper CI limit of the greater mean, ($\bar{Y}_2$) in this case, where d is understood to equal the difference between the two means. Equation 6 proves that when R_g is less than Delta, the maximum difference that is unimportant or can be dismissed on substantive grounds, then the standard 95% CI test for statistical equivalence is also satisfied. Said otherwise, the high side of the CI for the difference is also less than the high end of the Delta interval. One can visually determine statistical equivalence by graphing a bar whose height equals Delta, beginning at the lower limit of the lesser mean, as illustrated in Figure 2. Statistical equivalence obtains if the upper limit of the greater mean is less than or equal to the top of the Delta bar. The two means graphed in Figure 2 are statistically equivalent because the upper limit of the greater mean is below the top of the Delta bar (i.e., R_g is less than Delta).

Figure 2 can also be understood in terms of the maximum probable difference between the two means. All equal-length intervals within the CI have the same probability that the true value lies within them (see Interpreting CIs section for justification). The subinterval in which the mean falls is therefore no more likely to contain the true value than is any other interval. The maximum probable difference between the two means results when the lower mean, ($\bar{Y}_1$) in this case, is placed against the bottom of its inferential CI and the higher mean, ($\bar{Y}_2$) in this case, is placed against the top of its inferential CI. This maximum mean difference equals R_g. Statistical equivalence therefore is said to result when the maximum probable mean difference estimate provided by the inferential CIs is less than the amount that defines equivalence, or Delta. This is a very understandable and sensible definition of statistical equivalence.

What about the possibility that R_g equals Delta? It could be argued that this condition also satisfies statistical equivalence because the Delta amount is of no consequence and R_g is the same amount. However, the concept of a CI wholly contained within a band of

[6] Nickerson (2000, p. 255) extended Meehl's distinction between strong and weak uses of NHST by discriminating between NHST acceptance–support and rejection–support.

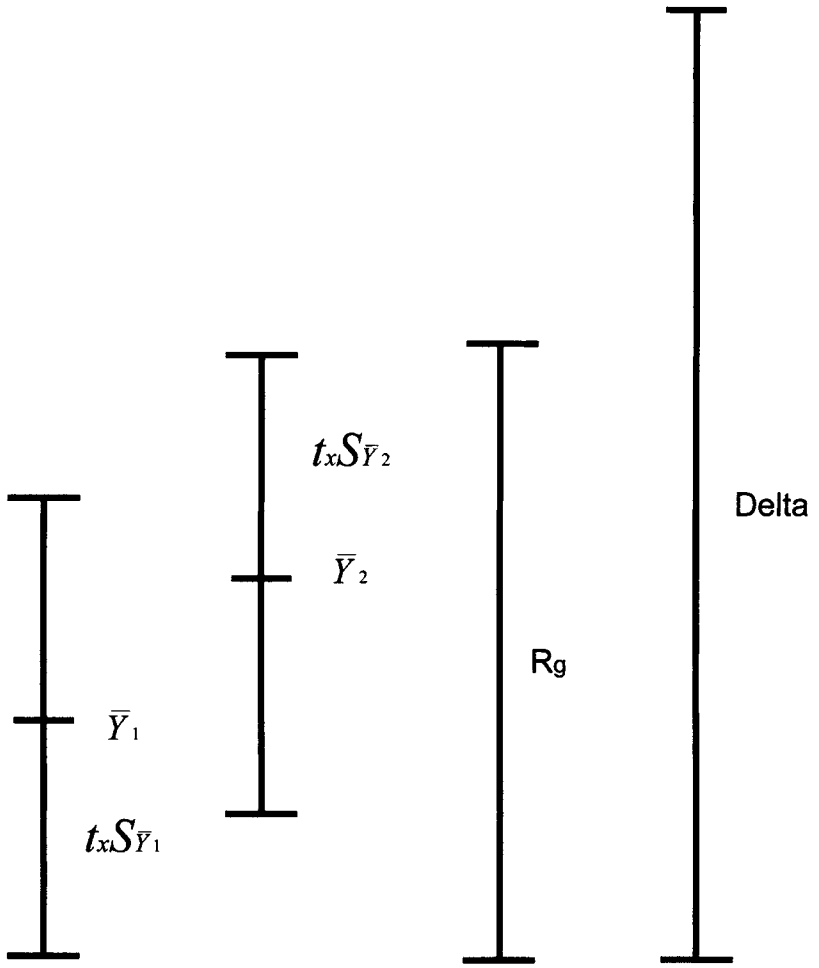

Figure 2. Two overlapping confidence intervals that are statistically equivalent because the confidence interval range (R_g) is less than the amount that substantively defines equivalence (Delta).

no consequence argues against accepting $R_g = \Delta$ as evidence of statistical equivalence. One can best avoid this matter by extending the precision of measurement. If we really mean that a difference of 3.00 is acceptable, then Delta should be set to something trivially greater, such as 3.01.

I now use the data for Groups B and C in Table 1 to illustrate the required calculations for statistical equivalence. The first step is to specify the maximum amount of difference (R_g) that one is willing to ignore in the name of equivalence on substantive grounds. For illustrative purposes, we set $\Delta = 3.0$ in this example. This means that the difference between the lower limit of the lesser mean and the upper limit of the greater mean (R_g) must be less than or equal to 3.0 for statistical equivalence to exist. We obtain standard errors for each group by dividing the standard deviation of each group by the square root of the number of subjects in each group. The E calculation in this case is as follows:

$$E = \frac{\sqrt{(0.7054)^2 + (0.5293)^2}}{0.7054 + 0.5293} = .7143.$$

The equal sample size of 20 subjects resulting in 19 degrees of freedom is associated with a 5% critical t value of 2.0930. This results in $t_x = .7143\ (2.0930) = 1.4950$ and the following inferential CIs based on standard errors taken from Table 1:

Group B: $75 \pm 1.4950\ (0.7054) = 73.945$ to 76.055

Group C: $76 \pm 1.4950\ (0.5293) = 75.209$ to 76.791.

The difference between the lower limit of the lesser mean of 73.945 and the upper limit of the greater

mean of 76.791 is $R_g = 2.846$, which is less than the stipulated Delta value of 3.0, and therefore we conclude that the means of these two groups are statistically equivalent. The p value for this statement is .05 because the critical t value associated with the 5% significance level was initially chosen before being reduced by the E factor to obtain the inferential CI.

Statistical Indeterminacy

The quantitative and graphic methods presented above test whether means are statistically different or equivalent. We are concerned here with the remaining outcome, in which means are neither statistically different nor equivalent. These cases are statistically indeterminate. Judgment must be suspended in such cases because statistical indeterminacy is not evidence for or against anything; it is not evidence of any kind. Additional data may result in positive evidence of statistical difference or equivalence, or indeterminacy may persist. It is expected that investigators will more frequently avail themselves of this alternative, because a conclusion of indeterminacy is less harsh than acceptance of the null hypothesis. This should reduce the frequency with which nonsignificant differences are termed a trend and interpreted with nearly as much force as a statistically significant difference.

Conclusions regarding statistical indeterminacy follow from conducting a test of statistical difference and a test of statistical equivalence. Hence, both tests should always be performed to guard against misinterpretation. The recommended approach facilitates conducting both tests.

Distribution Dependent

CIs do not circumvent distributional assumptions (cf. Steiger & Fouladi, 1997). Hence, it is important to test data for normality prior to constructing the above mentioned inferential CIs when one is testing for statistical difference and equivalence. Highly skewed data continue to present problems for the CI approach recommended here.

Power, Sample Size, and Test Reliability

The statistical power of the recommended procedures is identical to that of the standard t test because the recommended procedures are algebraically equivalent to the standard t test. Hence, procedures for determining sample size for statistical difference are the same as for the standard t test.

Statistical equivalence requires that the range (R_g) stemming from the lower inferential CI limit of the lesser mean to the upper inferential CI limit of the greater mean is less than Delta, the difference deemed inconsequential on substantive grounds. If the difference between the two means exceeds Delta, then statistical equivalence can never be demonstrated, regardless of how many subjects are studied. This fact can be used to detect pointless research proposals or doomed studies that never had a chance. If the difference between the two means is less than Delta, then it is always possible to shrink the inferential CIs so that they also fit within Delta, given enough subjects. The required sample size can be determined as follows. Subtract the difference between the two means from Delta to see what room is left. Subtract a trivial amount (e.g., .01) to enable R_g to be less than Delta (see above). Divide the remaining difference in half to determine how narrow each half-inferential CI must be. Recompute standard errors, E, and inferential CIs on the basis of various sample sizes until both CIs fit within Delta.

Measurement that is more reliable will further reduce uncertainty and thereby increase statistical power. Anastasi (1988, p. 133) showed that the standard error of measurement of a single score is a function of test reliability, as specified in Equation 7:

$$S_e = S_Y\sqrt{1 - r_{YY}}. \quad (7)$$

When reliability equals .75, the standard error of measurement is half of the sample standard deviation, and when reliability is .99, the standard error of measurement is one tenth of the sample standard deviation. Less variable single scores produce smaller group standard deviations and therefore smaller CIs. This increased measurement precision will result in greater statistical power.

Effect Size

CIs are centered about the obtained statistic rather than about zero and thereby call attention to effect size. Serlin (1993) correctly pointed out that confirmation and falsification depend directly on effect size. If data support a theory, then we want to know how large the effect size is. If data do not support a theory, then we want to know how close we came to obtaining support. Was theory slightly wrong, suggesting minor modification, or was theory largely incorrect, suggesting major revision or complete rejection? CIs provide the relevant information.

Failure to reject the null hypothesis is sometimes incorrectly interpreted as failure to replicate an earlier study (Krantz, 1999, p. 1374). If CIs overlap substantially with previously published intervals, then it is difficult to interpret one as a failure to replicate the other. Strong positive evidence of replication occurs when evidence of statistical equivalence is obtained. Strong negative evidence of replication occurs when evidence of statistical difference is obtained. Statistical indeterminacy is not evidence of any kind and therefore neither supports nor contradicts replication.

Cause–Effect Ratio

Abelson (1997b) pointed out that we rarely discuss effect size in the context of cause size. Small effect sizes can be quite interesting, especially when produced by a small cause size. Abelson (1997b) recommended that we calculate the ratio of these two quantities. In keeping with the convention for derivatives as change in y divided by change in x and the convention of y as the dependent variable and x as the independent variable, the cause–effect ratio should be effect size divided by cause size. Effect size can be the difference between the two means or the difference between two means divided by a common standard deviation or the standard deviation of the control group. Cause size can be drug dosage (mg/kg), stimulus intensity (dB for sound), number of treatment sessions, and so forth. This derivative-inspired cause–effect ratio can be used to compare results across studies that have manipulated a presumed cause. Greater similarity across studies might be found once cause size is controlled in this way.

Interpreting CIs

Abelson (1997a) warned that "under the Law of Diffusion of Idiocy, every foolish application of significance testing will beget a corresponding foolish practice for confidence limits" (p. 13). It is therefore crucial that the reader has a correct understanding of CIs. Perhaps a computational presentation is best understood. Imagine the following experiment. We begin with a normally distributed population containing a suitably large number of cases. The mean of the population is termed a parameter and is symbolized by μ. A sample of size N is randomly drawn from this population, and a 95% CI is constructed around it. We record "yes" if this CI includes the numerical value of μ and "no" if it does not. We repeatedly draw samples of the same size from this population a suitably large number of times, constructing a 95% CI and recording "yes" or "no" depending on whether the CI contains the numerical value of μ in each case. On average, we expect 950 "yes" and 50 "no" observations for every 1,000 replications.

A horseshoe example may also be helpful.[7] Howell (1997) stated,

> The parameter μ is not a variable—it does not jump around from experiment to experiment. Rather, μ is a constant, and the (confidence) *interval* is what varies from experiment to experiment. Thus, we can think of the parameter as a stake and the experimenter, in computing confidence limits, as tossing rings at it (CIs are shaped somewhat like horseshoes). Ninety-five percent of the time, a ring of specified width will encircle the parameter; 5% of the time, it will miss. A confidence statement is a statement of the probability that the ring has been on target; it is *not* a statement of the probability that the target (parameter) landed in the ring. (p. 204)

The CI can also be understood as containing all of the null hypotheses that cannot be rejected. No one null hypothesis is any more likely to be rejected than any other one. Hence, one cannot speak of any one position within the CI as any more likely to occur than any other. One can only talk about a parameter as within or without a CI, because one can only accept or reject the null hypothesis. No normal curve is superimposed on the CI. Likelihood is a Bayesian concept and is not relevant to classic NHST. Support for this position is found in Kirk (1982), who stated, "This confidence interval indicates values of the parameter μ that are consistent with the observed sample statistic. It also contains a range of values of μ_0 for which the null hypothesis is nonrejectable at the .05 level of significance" (p. 43). Winer et al. (1991) stated,

> The confidence statement, then, is a statement about the relative frequency with which limits, which vary from sample to sample, will enclose the fixed parameter. The statement C ($20 \leq \mu \leq 50$) = 0.95 means that if this experiment is repeated many times 95 percent of all limits computed would enclose the fixed value μ, while 5 percent would not include the parameter. (p. 47)

Clearly, no one value within the CI, reflecting a particular null hypothesis, is privileged as being more likely than any other. Therefore, all values within the CI are to be understood as equally likely.

---

[7] I thank Dr. Herman Friedman (personal communication) for bringing the horseshoe analogy to my attention.

No other valid interpretations for CIs exist within the classic NHST perspective. No further implications, assurances, or other meanings are to be ascribed to CIs.

Decision Rule Summary

Calculate inferential CIs by reducing the descriptive CIs by the E factor and draw one of the following three conclusions on graphical and/or numeric grounds:

1. There is statistical difference if the inferential CIs do not overlap, or
2. There is statistical equivalence if both inferential CIs, R_g, fit within the Delta (Δ) bound of indifference, or
3. There is statistical indeterminacy in all other cases.

Caveats

NHST abuses are viewed here as the result of misunderstandings regarding the null hypothesis. The goal of the present article is to remediate this human factors problem using an alternative NHST presentation that makes the null hypothesis implicit rather than explicit. This approach does not replace or avoid the null hypothesis. Rather, the recommended procedures focus on criteria for establishing statistical difference, equivalence, and indeterminacy.

No new statistical theory has been proposed. No claim to theoretical originality or advance of statistical theory has been made, with the possible exception that the reduced CIs introduced by Goldstein and Healy (1995) have been termed and treated here as inferential CIs in contrast with traditional descriptive CIs.

The recommended procedures are algebraically equivalent to standard NHST procedures and therefore impose the same evidential standard. Criticisms that the recommended procedures could be accomplished by more traditional methods miss the point that 75 years of traditional presentations have created rather than remedied the major human factors problem at hand. Only time will tell whether the recommended procedures actually repair this human factors problem.

References

Abelson, R. P. (1997a). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8,* 12–15.

Abelson, R. P. (1997b). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–141). Mahwah, NJ: Erlbaum.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Communications in Statistics—Theory and Methods, 12,* 2663–2692.

Andrews, H. P., Snee, R. D., & Sarner, M. H. (1980). Graphical display of means. *The American Statistician, 34,* 195–199.

Arbuthnot, J. (1710). An argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions of the Royal Society of London, 27,* 186–190.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57,* 289–300.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325–335.

Blackwelder, W. C. (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials, 3,* 345–353.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48,* 378–399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61,* 287–292.

Chow, S. L. (1998). Statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21,* 169–239.

Cohen, J. (1994). The world is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical tests in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology, 62,* 75–82.

Detsky, A. S., & Sackett, D. L. (1985). When was a 'negative' clinical trial big enough? How many patients you needed depends on what you found. *Archives of Internal Medicine, 145,* 709–712.

Dunnett, C. W., & Gent, M. (1977). Significance testing to establish equivalence between treatments with special reference to data in the form of 2 × 2 tables. *Biometrics, 33,* 593–602.

Estes, W. K. (1997a). On the communication of information

by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review, 4,* 330–341.

Estes, W. K. (1997b). Significance testing in psychological research: Some persisting issues. *Psychological Science, 8,* 18–20.

Fisher, R. A. (1925). *Statistical methods for research workers.* London: Oliver & Boyd.

Frick, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition, 23,* 132–138.

Gabriel, K. R. (1978). A simple method of multiple comparisons of means. *Journal of the American Statistical Association, 73,* 724–729.

Glass, G. V., & Hopkins, K. D. (1970). *Statistical methods in education and psychology* (3rd ed.). Boston: Allyn and Bacon.

Goldstein, H., & Healy, M. J. R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 158A,* Part 1, 175–177.

Guilford, J. P., & Fruchter, B. (1973). *Fundamental statistics in psychology and education* (5th ed.). New York: McGraw-Hill.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52, 15–24.*

Harlow, L. L. (1997). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 1–17). Mahwah, NJ: Erlbaum.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Harris, R. J. (1997a). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145–174). Mahwah, NJ: Erlbaum.

Harris, R. J. (1997b). Significance tests have their place. *Psychological Science, 8,* 8–11.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75,* 800–802.

Hochberg, Y., Weiss, G., & Hart, S. (1982). On graphical procedures for multiple comparisons. *Journal of the American Statistical Association, 77,* 767–772.

Hodges, J., & Lehmann, E. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B, 16,* 261–268.

Hogben, L. (1957). *Statistical theory.* London: Allen & Unwin.

Howell, D. C. (1997). *Statistical methods for psychology* (4th ed.). Belmont, CA: Duxbury Press.

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings* (pp. 29–33). Newbury Park, CA: Sage.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67,* 160–167.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparisonwise Type I error control. *Psychological Methods, 4,* 58–69.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Reveiw, 24,* 328–338.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44,* 1372–1381.

Loftus, G. R. (1991). On the tyranny of hypothesis testing in the social sciences. *Contemporary Psychology, 36,* 102–105.

Loftus, G. R. (1993). A picture is worth a thousand *p* values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers, 25,* 250–256.

Loftus, G. R. (1995). Data analysis as insight: Reply to Morrison and Weaver. *Behavior Research Methods, Instruments & Computers, 27,* 57–59.

Loftus, G. R. (1996). Psychology will be much better science when we change the way we analyse data. *Current Directions in Psychological Science, 5,* 161–171.

Loftus, G. R., & Mason, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490.

Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Makuch, R., & Simon, R. (1978). Sample size requirements for evaluating a conservative therapy. *Cancer Treatment Reports, 62,* 1037–1040.

Markus, K. A. (2001). The converse inequality argument against tests of statistical significance. *Psychological Methods, 6,* 147–160.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variation of box plots. *The American Statistician, 32,* 12–16.

McMan, J. C. (1995, August). *Statistical significance testing fantasies in introductory psychology textbooks.* Paper presented at the 103rd Annual Convention of the American Psychological Association, New York.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psy-

chology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Meehl, P. E. (1990a). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry, 1,* 108–141.

Meehl, P. E. (1990b). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244.

Metzler, C. M. (1979). Bioavailability—a problem in equivalence. *Biometrics, 30,* 309–317.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy.* Chicago: Aldine.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–115). Mahwah, NJ: Erlbaum.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Pearce, S. C. (1992). Introduction to Fisher (1925) statistical methods for research workers. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics, Volume II: Methodology and distributions* (pp. 59–65). New York: Springer-Verlag.

Rindskopf, D. M. (1997). Testing "small," but not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 319–332). Mahwah, NJ: Erlbaum.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significant tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Rosenthal, R. (1995). Progress in clinical psychology: Is there any? *Clinical Psychology: Science and Practice, 2,* 133–150.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis signficance test. *Psychological Bulletin, 57,* 416–428.

Scarr, S. (1997). Rules of evidence: A larger context for the statistical debate. *Psychological Science, 8,* 16–17.

Schmidt, F. (1996). APA Board of Scientific Affairs to study issue of significance testing, make recommendations. *The Score Newsletter, 19,* 1, 6.

Schmidt, F., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Selwyn, M. R., Dempster, A. P., & Hall, N. R. (1981). A Bayesian approach to bioequivalence for the 2 × 2 changeover design. *Biometrics, 37,* 11–21.

Selwyn, M. R., & Hall, N. R. (1984). On Bayesian methods for bioequivalence. *Biometrics, 40,* 1103–1108.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61,* 350–360.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Shaffer, J. P. (1986). Modified sequentially rejected multiple test procedures. *Journal of the American Statistical Association, 81,* 826–831.

Shaffer, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology, 46,* 561–584.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to special section exploring the pros and cons. *Psychological Sciences, 8,* 1–2.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Hillsdale, NJ: Erlbaum.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Science, 61,* 1340–1341.

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741–744.

Westlake, W. J. (1979). Statistical aspects of comparative bioavailability trials. *Biometrics, 35,* 273–280.

Westlake, W.J. (1981). Response to T. B. L. Kirkwood: Bioequivalence testing—A need to rethink. *Biometrics, 37,* 589–594.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist, 54,* 594–604.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons with examples from state-to-state differences in educational achievement. *Journal of Education and Behavioral Statistics, 24,* 42–69.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.) New York: McGraw-Hill.

(Appendix follows)

Appendix

Raw Simulated Data for Computational Examples in Tables 1 and 2

A	B	C	D	E	F
81.6677	70.7683	75.7627	60.7260	58.6962	47.9228
67.5400	80.3705	69.6446	63.2949	45.9228	64.6962
62.5488	78.7280	75.7906	69.8218	54.5076	58.5076
62.1265	72.4568	77.1663	75.6620	52.8336	56.8336
47.9714	70.7176	74.2195	78.7183	51.9021	63.3030
66.7972	75.8592	75.7018	63.8045	59.3030	55.9021
48.9645	75.2776	77.5868	56.7752	51.0099	55.0099
51.8507	71.5389	76.2926	75.9659	49.3595	53.3595
77.0004	77.7434	77.2740	77.9042	44.2099	48.2099
81.0220	74.0183	77.7969	53.7597	53.3770	57.3770
65.7478	72.8827	70.9844	75.2164	40.1522	44.1522
77.0959	78.9173	76.2522	67.3436	43.7642	47.7642
92.7205	76.6139	76.0068	64.9018	53.7458	57.7458
73.4298	74.6258	77.2068	71.7976	54.1364	58.1364
81.3882	77.0918	76.9766	82.5448	54.5431	58.5431
56.5072	79.1296	78.4739	54.3386	47.8768	51.8768
60.0742	73.8517	73.6080	82.4223	50.6031	54.6031
60.1151	69.6584	76.4078	80.2158	45.8731	49.8731
69.7974	72.8166	77.3037	88.8500	45.3379	49.3379
75.6352	76.9331	79.5436	74.0128	42.8460	46.8460
			63.2426		
			78.1053		
			91.2740		
			49.6895		
			74.6134		

Received April 27, 2000
Revision received June 22, 2001
Accepted July 26, 2001 ■