Using Significance Tests to Evaluate Equivalence Between Two Experimental Groups

James L. Rogers, Kenneth I. Howard, and John T. Vessey

Equivalency testing, a statistical method often used in biostatistics to determine the equivalence of 2 experimental drugs, is introduced to social scientists. Examples of equivalency testing are offered, and the usefulness of the method to the social scientist is discussed.

Although the central limit theorem was developed to allow for the estimation of confidence bounds around an observed mean (see Adams, 1974, for a fascinating presentation), its major application in empirical science has been to test whether the absolute difference between two means is greater than zero. However, there has been a growing dissatisfaction with traditional tests of the null hypothesis, in which the difference between two population means is precisely zero. Somehow, the testing of a hypothesis of "no difference" has resulted in the cognitive illusion that the investigator did not actively choose this as a plausible alternative hypothesis—that the null hypothesis was just a given of nature. However, it has been long recognized that with very large sample sizes, this null hypothesis will be rejected in almost all cases, resulting in statistically significant differences that are substantively trivial. In response to this state of affairs, and after establishing the statistical reliability of results, investigators have turned to estimates of the "amount of variance accounted for" or effect sizes (ESs) to evaluate the substantive significance of their findings. With the recent popularity of power analysis, investigators now design their studies in such a way that statistical analyses will be relevant to preselected differences (e.g., small, moderate, or large ESs) of presumed substantive import. This has resulted in a more complex interplay between hypothesis testing and statistical analyses: one in which investigators are asked to select a meaningful difference before executing a study.

Dissatisfaction with the traditional null hypothesis has also emerged in an area of research in which the aim is not to establish the superiority of one treatment or method over another, but rather to establish equality between the two methods. This type of research involves the testing of treatment innovations to determine if a new method achieves an equally effective outcome as the standard method but perhaps at a lower cost or greater convenience. For example, an investigator may hope to show that group therapy is as effective as individual therapy or that a less expensive antidepressant medication works as well as a more costly one. In these cases, the question is "Is there a more efficient way to achieve the same result?"

Common to the increasing interest of social scientists in power analysis, ES specification, and the equality of treatments is the realization that the objective is often to determine whether mean values are "equivalent" rather than "different." Though largely unfamiliar to social scientists, formal statistical tests of equivalence have been evolving over the past 20 years. At present, equivalency tests fall into three general categories: the *confidence interval approach,* developed by Westlake (1981) and presented in this article; the *nonequivalence null hypothesis approach,* developed by Anderson and Hauck (1983), which uses an approximation to a noncentral t distribution to calculate the p level of the test; and *Bayesian methods,* developed by Selwyn, Dempster, and Hall (1981) and Selwyn and Hall (1984). The first two methods are the most attractive because they require the fewest arbitrary decisions (Westlake, 1988). Comparisons of the two methods have revealed that the confidence interval method is conservative (i.e., the actual Type I error rate is equal to or less than the stated Type I error rate) and the Hauck and Anderson method can be liberal (i.e., the actual Type I error rate could be higher than the stated Type I error rate [Anderson and Hauck, 1983]).

In this article, the method (Westlake, 1981) typically used by biostatisticians to determine if two drugs have an equivalent impact (Hauck & Anderson, 1986; Makuch & Simon, 1978; Westlake, 1988) is introduced to social scientists. Whereas the purpose of a traditional hypothesis test is to determine whether two groups differ from one another, this procedure is used to determine whether two groups are sufficiently near each other to be considered equivalent. Equivalency testing is appropriate when the investigator is able to specify a small, nonzero difference between two treatments that would serve to define an "equivalence interval" around a difference of zero (e.g., ±10%). Any difference small enough to fall within that equivalence interval would be considered clinically and/or practically unimportant.

James L. Rogers, Department of Psychology, Wheaton College; Kenneth I. Howard and John T. Vessey (now at the Office of Population Affairs, U.S. Department of Health and Human Services, Washington, DC), Department of Psychology, Northwestern University.

This work was partially supported by Research Grant R 01 MH42901 from the National Institute of Mental Health. We thank Amy Miller for her technical assistance.

Correspondence concerning this article should be addressed to James L. Rogers, Department of Psychology, Wheaton College, Wheaton, Illinois 60187.

Psychological Bulletin, 1993, Vol. 113, No. 3, 553-565
 0033-2909/93/$3.00

Equivalency testing is straightforward, using concepts highly familiar to social scientists. Both Type I and Type II error rates are controlled. There is a null hypothesis asserting that the difference between two groups is at least as large as the one specified by the investigator, and there is an alternative hypothesis asserting that the difference between two groups is smaller than the specified one. As in traditional hypothesis testing, the goal of the investigator is to reject the null hypothesis and accept the alternative hypothesis. In the discussion that follows, we describe equivalency testing, provide formulas to establish equivalency between either two means or two proportions, and offer a number of illustrative examples. We also examine sample size estimation procedures that allow equivalency testing at a designated level of power. Finally, we discuss the usefulness of equivalency testing in the social sciences.

Method

Procedure

Equivalency testing is accomplished in two steps: The investigator first defines equivalency and then performs two simultaneous one-sided hypothesis tests.

Defining equivalency. An a priori decision must be made concerning the minimum difference between two groups that would be important enough to make the groups nonequivalent. The investigator will typically consider two means (or proportions) equivalent if they differ by less than some delta in both a negative (δ_1) and positive (δ_2) direction. If a greater difference in one direction than the other is allowed, δ_1 and δ_2 should be individually defined; otherwise, $\delta_2 = -\delta_1$.

The equivalence definition will depend on the substantive issue under consideration. Equivalence between an experimental group and a control group might be a difference of less than 20% of the control group mean (if the metric is appropriate), a difference of less than 20% of the pooled standard deviation, or a difference less than the minimum value considered to be substantively important. In certain instances, more than one equivalency definition might be specified, each tailored to a particular application or experimental perspective. Test results would then be reported for each definition.

Two simultaneous one-sided tests. Two one-sided hypothesis tests must be performed. Figures 1A and 1B illustrate the test procedure, whereas Table 1 presents the required formulas to establish equivalency between two means or two proportions. Test 1 seeks to reject a null hypothesis asserting that the difference between two means (or proportions) is less than or equal to the smaller delta (δ_1). Test 2 seeks to reject a null hypothesis asserting that the difference is greater than or equal to the larger delta (δ_2).

Because δ_1 and δ_2 are the minimum differences (in a negative and positive direction) that would make a difference, the investigator's goal is to demonstrate statistically that an observed difference between two means, $M_1 - M_2$, is too large to have come from a distribution with mean of δ_1 (Test 1) and simultaneously too small to have come from a distribution with mean δ_2 (Test 2).

The logic behind the test is that if $M_1 - M_2$ is shown to have come from a distribution simultaneously to the right of δ_1 and to the left of δ_2, the investigator can conclude that the distribution it came from is somewhere in the middle, with true difference $\mu_1 - \mu_2$ less than the minimum difference of importance that was determined by the investigator.

Note that to establish equivalency the investigator must reject both one-sided null hypotheses; however, to do so, only one test is required. Consider the case in which the observed difference, $M_1 - M_2$, is of *unequal distance* between δ_1 and δ_2. Of the two one-sided tests that must be rejected, one test evidences a shorter distance between its observed value ($M_1 - M_2$) and its null value for delta (δ_1 or δ_2). Choosing the one-sided test having the shorter distance between $M_1 - M_2$ and delta (either δ_1 or δ_2) will yield the smaller test statistic and consequently the larger p value of the two possible tests.[1] Because this test has the larger p value, it will be the least likely to show equivalence. However, if the test with the larger p value is rejected, it follows that the remaining one-sided test, which will necessarily evidence a smaller p value, need not be performed as it will always be rejected as well. On the other hand, if the test in question (the largest p value) does not result in the rejection of its null value, it will still be unnecessary to perform the second test because both tests must be rejected to conclude that $\mu_1 - \mu_2$ falls within the equivalency interval. In other words, there is never a case when the test of the larger difference between $M_1 - M_2$ and its delta (δ_1 or δ_2) will need to be conducted.[2] Finally, this statement is true even if $M_1 - M_2$ falls exactly between δ_1 and δ_2. Here, *the conclusion for one test will be identical to that of the other.*[3] Both will have the same p value (i.e., both test statistics will have the same absolute value). Thus, only one of the statistical tests needs to be done.

It follows from the discussion above that for an equivalency test, the probability of a Type I error is equal to the alpha (α_1 or α_2) selected by the investigator to evaluate the one-sided test evidencing the greatest p value. This test—the one actually performed—predicts perfectly the outcome of the second test. Because results of the two tests are completely dependent, the Type I error rate does not need to be adjusted to account for both tests. Thus, the alpha for an equivalency test is the value given to the one-sided test actually conducted (α_1 or α_2), that is, the alpha corresponding to the test evidencing the largest p value.[4]

The contrast between traditional and equivalence procedures provides another view of Type I error in equivalency testing. When conducting a traditional two-tailed test, the experimenter rejects the null hypothesis if either of the two test statistics is significant.[5] Because in a two-tailed test either test statistic being significant by chance would lead to a Type I error, the experimenter must add the Type I error probabilities for each of the two test statistics to calculate the overall probability of making a Type I error. However, in an equivalency test,

---

[1] Let $|(M_1 - M_2) - \delta_1| > |(M_1 - M_2) - \delta_2|$. Then $z_1 = |(M_1 - M_2) - \delta_1|/s_{M_1-M_2} > z_2 = |(M_1 - M_2) - \delta_2|/s_{M_1-M_2}$. Therefore, $p(z_2) > p(z_1)$. Similarly, if $|(M_1 - M_2) - \delta_1| < |(M_1 - M_2) - \delta_2|$, then $p(z_1) > p(z_2)$.

[2] If $p(z_2) > p(z_1)$ and $p(z_2)$ leads to rejection of the null, so will $p(z_1)$. By the same logic, if $p(z_1) > p(z_2)$ and $p(z_1)$ leads to rejection of the null, so will z_2.

[3] Let $|(M_1 - M_2) - \delta_1| = |(M_1 - M_2) - \delta_2|$, then $z_1 = z_2$, so $p(z_1) = p(z_2)$. Both tests lead to the same conclusion. Thus, only one test need be performed.

[4] It is sometimes difficult for those newly exposed to equivalency testing to understand why overall α is not adjusted to account for the fact that two tests are performed. However, it should be clear from the discussion here that no adjustment is necessary. If a Type I error is made, it is because the distribution that $M_1 - M_2$ comes from is actually to the left of δ_1 or to the right of δ_2, but it cannot be both. Therefore, because we do not know on which side the actual distribution lies, the investigator takes a worst-case scenario approach and chooses to examine the larger of the p values from the two tests (p_1 and p_2 in Figure 1B). This corresponds to choosing the test that has a higher probability of a Type I error. If this larger p value is less than α, then the investigator will assume (but cannot prove) that a Type I error was not made. Because the two tests constituting the equivalency procedure are conducted simultaneously—that is, the larger test statistic is conditioned on the smaller—only one decision is made, and that decision is unidirectional. Namely, a single difference is judged to be either acceptably small or unacceptably large.

[5] A traditional two-tailed test is in fact two tests, one for each tail.

both test statistics must be significant to lead an experimenter to reject the null hypothesis. Because the first test statistic and the second test statistic have to be significant by chance for a Type I error to be made, the experimenter in this case would multiply the Type I error probabilities of each test statistic to calculate the overall probability of making a Type I error. However, as noted above, the probability of the larger (absolute value) of the two test statistics being significant, given that the smaller of the two is significant, equals 1. That is, the overall Type I error probability is the Type 1 error probability of the smaller test statistic multiplied by the conditional Type I error probability of the larger statistic, which will always be 1. Therefore, the overall probability of making a Type I error in an equivalency test is simply the Type I error associated with the smaller of the two test statistics.

Confidence intervals. Rather than conducting two one-sided tests, as described above, a confidence interval may be constructed. Equivalence is concluded if the confidence interval is contained within the equivalence interval. (E.g., a symmetrical equivalence interval would be the interval bounded by δ_1 and $\delta_2 = -\delta_1$.) Note that the equivalency confidence interval should be expressed at the $1 - 2\alpha$ level of certainty rather than at the customary $1 - \alpha$ level.[6] The $1 - 2\alpha$ confidence interval will fall within the equivalence interval when both one-sided tests are simultaneously rejected, thereby leading to the rejection of the null with α probability of a Type I error.

Sample size formulas. The sample size per group (n_T) required for a means test can be obtained by either Formula 1 or 2 (see below) if delta is symmetrical with regard to direction ($\delta_1 = -\delta_2$) and $\mu_{Ha} = 0$.[7] Otherwise, the larger of the two sample sizes should be used.

$$n_{T1} = \frac{2s^2_{\text{pooled}}(z_\alpha + z_{\beta/2})^2}{(\mu_{Ha\alpha} - \delta_1)^2} \text{ for Test 1.} \quad (1)$$

$$n_{T2} = \frac{2s^2_{\text{pooled}}(z_\alpha + z_{\beta/2})^2}{(\delta_2 - \mu_{Ha\alpha})^2} \text{ for Test 2.} \quad (2)$$

Nearly the same formulas are applicable when proportions are tested.

$$n_{T1} = \frac{[\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \beta_2)](z_\alpha + z_{\beta/2})^2}{[(p_1 - p_2) - \delta_1]^2} \text{ for Test 1.}$$

$$n_{T2} = \frac{[\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)](z_\alpha + z_{\beta/2})^2}{[\delta_2 - (p_1 - p_2)]^2} \text{ for Test 2.}$$

Finally, in many instances, the investigator will set $\mu_1 - \mu_2$ (or $p_1 - p_2$) to zero for the purpose of sample size estimation. On the other hand, if the investigator believes that a difference does in fact exist, but wants to test to determine if the difference is small enough to fall within a defined equivalency interval, the investigator need only set μ_{Ha} equal to the expected difference. The required sample size is then estimated as described above.

Examples

The following examples have been selected to provide both computational and substantive illustrations of equivalency testing. No criticism of the analyses provided in the original publications is intended. The studies we have selected and the equivalence intervals we apply make a computational point aimed at disclosing a range of situations that may arise when conducting equivalency tests. Investigators who are experts in their individual areas of research will need to determine when equivalency testing is useful and to define meaningful equivalence intervals relative to the substantive issues at hand, just as they now must define meaningful levels of alpha and power or meaningful ESs.

Example 1: MMPI Similarities

Cannon, Bell, Fowler, Penk, and Finkelstein (1990) compared alcoholic and drug-abusive subjects on the Minnesota Multiphasic Personality Inventory (MMPI). We reexamine comparisons between 207 subjects diagnosed as alcohol dependent and 49 subjects diagnosed as drug (but not alcohol) dependent. In addition to significance levels for traditional z values, Table 2 shows the results of an equivalency procedure (two one-sided tests) to determine whether the mean MMPI profile scores of drug-addicted subjects were within 10% of the mean MMPI scores of alcoholic subjects. We consider a difference of 10% or less on the MMPI to be clinically trivial.

To illustrate the computational aspects of equivalency testing, we arbitrarily selected the Masculinity–Femininity (*Mf*) scale (eighth row in Table 2). The same procedure applies to all other scales.

The information required to conduct equivalency testing, found in Columns 2–5 of Table 2, includes the means, standard deviations, and sample sizes of the two groups being compared. For the *Mf* scale, these values are $M_1 = 59.2$, $SD = 9.5$, and $n = 207$ for the alcohol group and $M_2 = 61.4$, $SD = 10.9$, and $n = 49$ for the drug group. Because the equivalence interval is defined as ±10% of the alcohol group mean, we calculate that $\delta_1 = -10\% \times 59.2 = -5.92$ and $\delta_2 = 10\% \times 59.2 = +5.92$, or simply that the equivalency interval is ±5.92. The obtained difference between the two group means ($M_1 - M_2$) is −2.2 and, using the formula provided in Table 1, has the following standard error:[8]

---

[6] That the $1 - 2\alpha$ rather than the $1 - \alpha$ confidence interval should be used is apparent when the rejection regions for the simultaneous one-sided tests are considered. We have $(M_1 - M_2) - \delta_1/s_{M_1-M_2} > z_\alpha$ and $(M_1 - M_2) - \delta_2 < -z_\alpha$. So $\delta_1 - (M_1 - M_2)/s_{M_1-M_2} < -z_\alpha$ and $z_\alpha < \delta_2 - (M_1 - M_2)/s_{M_1-M_2}$. However, because $-z_\alpha < z_\alpha$, the following inequality is apparent: $\delta_1 < (M_1 - M_2) - z_\alpha s_{M_1-M_2} < (M_1 - M_2) + z_\alpha s_{M_1-M_2} < \delta_2$. That is, the $1 - 2\alpha$ confidence interval, namely, $(M_1 - M_2) \pm z_\alpha s_{M_1-M_2}$, will always fall within an equivalence interval bounded by δ_1 and δ_2 if both one-sided tests are simultaneously rejected.

[7] From Figure 1A it is readily seen that $-z_\alpha = [(M_1 - M_2) - \delta_2]/s_{M_1-M_2}$ and $-z_{\beta/2} = [(M_1 - M_2) - \mu_{Ha}]/s_{M_1-M_2}$. Here, z_α defines the probability of a Type I error, $z_{\beta/2}$ defines one half the probability of a Type II error (1 − power), $M_1 - M_2$ is an unknown difference score corresponding to both $-z_\alpha$ and $z_{\beta/2}$, and μ_{Ha} is expected $\mu_1 - \mu_2$ under the alternative hypothesis. If we assume equal sample sizes ($n_1 = n_2 = n$) and substitute for $M_1 - M_2$, we see that for Test 2 $-z_\alpha = [(z_{\beta/2}s_{M_2-M_2} + \mu_{Ha}) - \delta_2]/s_{M_1-M_2}$, where $s_{M_1-M_2} = s_{\text{pooled}}\sqrt{2/n}$. Thus, $n_{T2} = 2s^2_{\text{pooled}}(z_\alpha + z_{\beta/2})^2/(\delta_2 - \mu_{Ha})^2$ for Test 2, where n_{T2} is the minimal sample size per group required for a given α and β. Likewise, $n_{T1} = 2s^2_{\text{pooled}}(z_\alpha + z_{\beta/2})^2/(\mu_{Ha} - \delta_1)^2$ for Test 1. If delta is symmetrical with regard to direction ($\delta_1 = -\delta_2$) and $\mu_{Ha} = 0$, then either sample size formula may be used (i.e., $n_{T1} = n_{T2}$). Otherwise, the larger of the two sample sizes should be used. Nearly the same formulas are applicable when proportions are tested, $n_{T1} = [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)](z_\alpha + z_{\beta/2})^2/[(p_1 - p_2) - \delta_1]^2$ and $n_{T2} = [\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)](z_\alpha + z_{\beta/2})^2/[\delta_2 - (p_1 - p_2)]^2$.

[8] We are grateful to an anonymous reviewer for the following insight. Throughout this article, when delta has been defined as a percentage of the control group mean, the resulting delta has been treated as a constant with no variance. This is common in practice as there is seldom a compelling reason to view delta as, itself, a stochastic value. On the other hand, if one wishes to incorporate into the definition of delta the variability inherent in the control group mean, thereby treating delta

$$s_{M_1-M_2} = \left\{\left[\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right]\left[\frac{1}{n_1}+\frac{1}{n_2}\right]\right\}^{1/2}$$
$$= \left\{\left[\frac{(207-1)(9.5)^2 + (49-1)(10.9)^2}{207+49-2}\right]\left[\frac{1}{207}+\frac{1}{49}\right]\right\}^{1/2}$$
$$= 1.554.$$

The traditional z test yields an obtained test statistic value of -1.416 with an associated p value of .078. That is, $z = -2.2/1.554 = -1.416$, $p = .078$. Alternatively, one could compute a 95% confidence interval, $M_1 - M_2 \pm (z_\alpha)(s_{M_1-M_2}) = -2.2 \pm (1.96)(1.554)$, or -5.245 to 0.845. Note that zero falls within the interval. Having failed to obtain a statistical significance by the traditional test, an investigator might suspect that there is not an important difference between the two groups, although direct evidence is lacking. Indeed, a p value of .078 might be interpreted as "borderline significance" by some investigators. Therefore, using the assumption that a difference of 10% of the alcohol group mean is important, an equivalency test is conducted.

At this point, the investigator either refers to the larger p value of the two one-sided tests described in Table 1 or determines whether a 90% confidence interval falls into the equivalence interval. The two one-sided tests are as follows.

$$z_1 = \frac{(59.2 - 61.4) - (-5.92)}{1.554} = 2.394,\ p = .008, \text{ and}$$

$$z_2 = \frac{(59.2 - 61.4) - (5.92)}{1.554} = -5.226,\ p = .000.$$

The larger p value of .008 allows the null hypothesis of nonequivalence to be rejected; that is, the smaller obtained z value of 2.394 exceeds the critical value ($z_{0.05} = 1.645$). Therefore, an investigator may conclude that the difference between the alcohol and drug conditions is within 10% of the alcohol group mean. In doing so, the investigator runs a Type I error risk of .05. Had the rejection region been as extreme as all z values of 2.394 or greater, the Type I error risk would have been equal to the p value of .008.

Clearly, comparing a 90% confidence interval with the equivalence interval results in the same conclusion. In Table 2, we see that the lower confidence limit = (59.2 − 61.4) − (1.645)(1.554) = −4.756 and the upper confidence limit = (59.2 − 61.4) + (1.645)(1.554) = 0.356. Thus, the 90% confidence interval (−4.756 to 0.356) is contained within the equivalency interval (±5.92), and we conclude, as before, that the two conditions are equivalent using the 10% criterion. Note that the confidence limits might just as easily be reexpressed as a percentage of the alcohol group mean by dividing each limit by $M_1 = 59.2$. If this were done, the confidence interval, stated as a percentage of the alcohol group mean, would be 8.03% (4.756/59.2) to 0.60% (0.356/59.2). As expected, this interval (8.03% to 0.60%) falls within ±10%, allowing equivalency to be concluded.

To graphically display the equivalency results for each of the MMPI scales, 90% and 95% confidence intervals, expressed as a percentage of the alcohol group mean, are plotted in Figure 2. The outer tick marks reflect the 95% interval (the traditional test) and the inner tick marks reflect the 90% interval (the equivalency test). If on visual inspection the 90% interval falls within the equivalence band (±10%), one may conclude equivalence with a 5% risk of Type I error. Also, if on visual inspection the 95% interval excludes zero (0%), the traditional hypothesis test of no difference may be rejected with a 5% risk of a Type I error.

It is informative to consider the results found in Figure 2. To assist in the interpretation of Figure 2, the MMPI scales have been grouped into four categories on the basis of the outcome of both the traditional test and equivalency test. The alcohol and drug group comparisons for Depression (*D*), Psychopathic Deviate (*Pd*), Paranoia (*Pa*), Hypomania (*Ma*), and Social Introversion (*Si*) were statistically different by the traditional test and failed to obtain statistical equivalence by the two one-sided equivalency procedure. (The 95% confidence intervals do not include zero, and the 90% confidence intervals do not lie completely within the preset equivalency interval.) Thus, the data strongly suggest that *D*, *Pd*, *Pa*, *Ma*, and *Si* differ in a clinically important fashion between alcoholic and other drug-dependent subjects.

Because the 95% confidence intervals for Lie, Frequency, Hysteria, *Mf*, and Psychasthenia include zero and the 90% confidence intervals lie completely within the preset equivalence bounds, we conclude that differences within these five scales fail to reach statistical significance by the traditional test and that the two comparison groups were statistically equivalent. That is, these five scales exhibit no clinically important differences between alcoholic and other drug-dependent subjects.

The correction scale was found to be statistically different across groups by the traditional test as well as statistically equivalent by the equivalency test. Although statistically different across groups, this difference is clinically unimportant.

Finally, differences between groups on the Hypochondriasis and Schizophrenia scales were not statistically significant either by the traditional test or by the equivalency test. The variability in these scales was too great to allow an accurate appraisal given the sample size used in this study.

Example 2: Equivalency in Meta-Analysis

Robinson, Berman, and Neimeyer (1990) used meta-analysis to compare the efficacy of different types of therapies in the treatment of depression. Using various techniques, Robinson et al. calculated an average ES that contrasted each of several therapeutic approaches—cognitive versus behavioral, psychotherapy versus drug therapy, and so on—for studies judged not to suffer from investigator allegiance to any particular treatment.

itself as a random variable, then the standard errors used for each of the two one-sided tests would be derived as follows: Let a equal the percentage of the control mean used to define delta. Then the difference to be converted to the Test 1 Z score can be expressed as $(M_1 - M_2) - (-aM_1) = M_1 + aM_1 - M_2 = (1 + a)M_1 - M_2$. The variance of this random variable is derived as follows: $Var[(1 + a)M_1 - M_2] = (1 + a)^2 VarM_1 + VarM_2 = (1 + a)^2(s_1^2/n_1) + (s_2^2/n_2)$. Substituting the pooled variance for s_1^2 and s_2^2, the standard error used to compute the Test 1 Z score is seen to be $\{s^2_{pooled}[(1 + a)^2/n_1) + (1/n_2)]\}^{1/2}$. Similarly, the standard error used to compute the Test 2 Z score would be $\{s^2_{pooled}[(1 - a)^2/n_1) + (1/n_2)]\}^{1/2}$. Here, $s^2_{pooled} = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$.

Of concern here is whether these ESs were less than ±0.20, an ES value classified as "small" by Cohen (1977) and considered by us to reflect little, if any, clinical relevance.

Table 3 shows the mean ESs in question along with standard errors provided by Robinson et al. (1990). A z statistic, corresponding p value, and appropriate confidence interval (95% for the traditional procedure and 90% for the equivalency procedure) have been provided for each of the 12 comparisons. The traditional p values were found by dividing each mean ES by its corresponding standard error to obtain a z statistic ($z = ES/SE$), then converting to a p value. The p value tabled for the equivalency test is the larger value found for the two one-sided tests, $z_1 = (ES - 0.20)/SE$ or $z_2 = (ES + 0.20)/SE$. The confidence intervals were obtained by adding or subtracting from each effect size either $1.96 \times SE$ (traditional) or $1.645 \times SE$ (equivalence). For example, we obtain the following calculations for the cognitive versus behavioral contrast.

Traditional z: $z = ES/SE = 0.12/0.09 = 1.333$, $p = .091$.

Traditional confidence interval:

$$ES \pm (z_{\alpha/2})(SE) = 0.12 \pm (1.96)(0.09), \text{ or } -0.056 \text{ to } 0.296.$$

Equivalence z: $z_1 = (ES + 0.20)/SE = (0.12 + 0.20)/0.09 = 3.556$, $p = .000$.

$z_2 = (ES - 0.20)/SE = (0.12 - 0.20)/0.09 = -0.889$, $p = .187$.

So we table the larger p value of 0.187.

Equivalence confidence interval:

$$ES \pm (z_{\alpha})(SE) = 0.12 \pm (1.645)(0.09), \text{ or } -0.028 \text{ to } 0.268.$$

The results above indicate that the contrast ES for cognitive versus behavioral therapy (ES = 0.12) is neither statistically different from zero ($p = .091$, and confidence interval includes zero) nor statistically equivalent ($p = .187$, and confidence interval falls outside ±0.20).

The information in Table 3 is of considerable interest. The analysis shows psychotherapy to be equivalent to the four treatments with which it is compared (drug therapy, combination drug, tricyclics, and combination tricyclics). Variability relative to ES is too large to determine either a difference or an equivalency in efficacy for the remaining eight comparisons: cognitive versus behavioral, cognitive versus cognitive–behavioral, behavioral versus cognitive–behavioral, cognitive versus general verbal, behavioral versus general verbal, cognitive–behavioral versus general verbal, cognitive versus drug therapy, and combination (tricyclics) vs. tricyclics. The results of the traditional test and the equivalency test are displayed in Figure 3.

Example 3: Assessing Baseline Equivalence

Zabin, Hirsch, and Boscia (1990) compared three groups of inner-city Black adolescent women: those who (a) had negative pregnancy test results, (b) were pregnant and carried to term, and (c) terminated their pregnancy by induced abortion. Zabin, Hirsch, and Emerson (1989) and others (e.g., Adler et al., 1990) have presented various conclusions that involve psychosocial profile comparisons between birth and abortion groups at 1 and 2 years following the pregnancy resolution decision. Because Zabin et al. (1990) did not adjust for baseline differences, the presumption is that the abortion and birth groups are equivalent on important baseline parameters.

Table 4 shows the percentage presence of 27 baseline characteristics in women who carried to term or women who aborted, along with the outcomes of both the traditional and the equivalency hypothesis testing procedures. Baseline equivalency was evaluated by determining whether the birth group mean was within 20% of the abortion group mean. In effect, this criterion implies that the baseline parameters should be within 20% of each other before any attempt is made to explain 1- and 2-year differences that might emerge on the basis of the pregnancy resolution decision.

Table 4 presents proportions, differences between proportions, standard errors, and the results of the traditional test and the equivalency test. Using the formulas in Table 1, the following calculations may be verified for the baseline variable "ever repeated grade." The same procedure applies to all the baseline variables.

Standard error:

$$s_{\hat{p}_1-\hat{p}_2} = \left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]^{1/2} = \left[\frac{(.343)(1-.343)}{141} + \frac{(.505)(1-.505)}{93}\right]^{1/2} = 0.065.$$

Traditional z:

$$z = (\hat{p}_1 - \hat{p}_2)/SE = -.162/0.065 = -2.474, \; p = 0.007.$$

Traditional confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}(SE) = -.162 \pm (1.96)(0.065), \text{ or } -.290 \text{ to } -.034.$$

Equivalence z:

Where $\delta_1 = -20\% \times 0.343 = -0.069$,

$$z_1 = (\hat{p}_1 - \hat{p}_2) - \delta_1/SE = (-0.162 + 0.069)/0.065 = -1.427, \; p = .923.$$

Where $\delta_2 = 20\% \times 0.343 = 0.069$,

$$z_2 = (\hat{p}_1 - \hat{p}_2) - \delta_2/SE = (-0.162 - 0.069)/0.065 = -3.522, \; p = .000.$$

Equivalence confidence interval: $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha}(SE) = -0.162 \pm (1.645)(0.065)$, or $-.270$ to $-.054$.

The difference between the proportion of abortion subjects and the proportion of carrying-to-term subjects who had repeated at least one grade (34.3% vs. 50.5%, respectively) is a statistically significant difference that is not small enough to be considered statistically equivalent. That is, the traditional z is greater than $z_{0.025} = 1.96$ (the 95% confidence interval does not contain zero), and the smaller equivalence z (with the higher p value) is not larger than $z_{0.05} = 1.645$. This means that the 90%

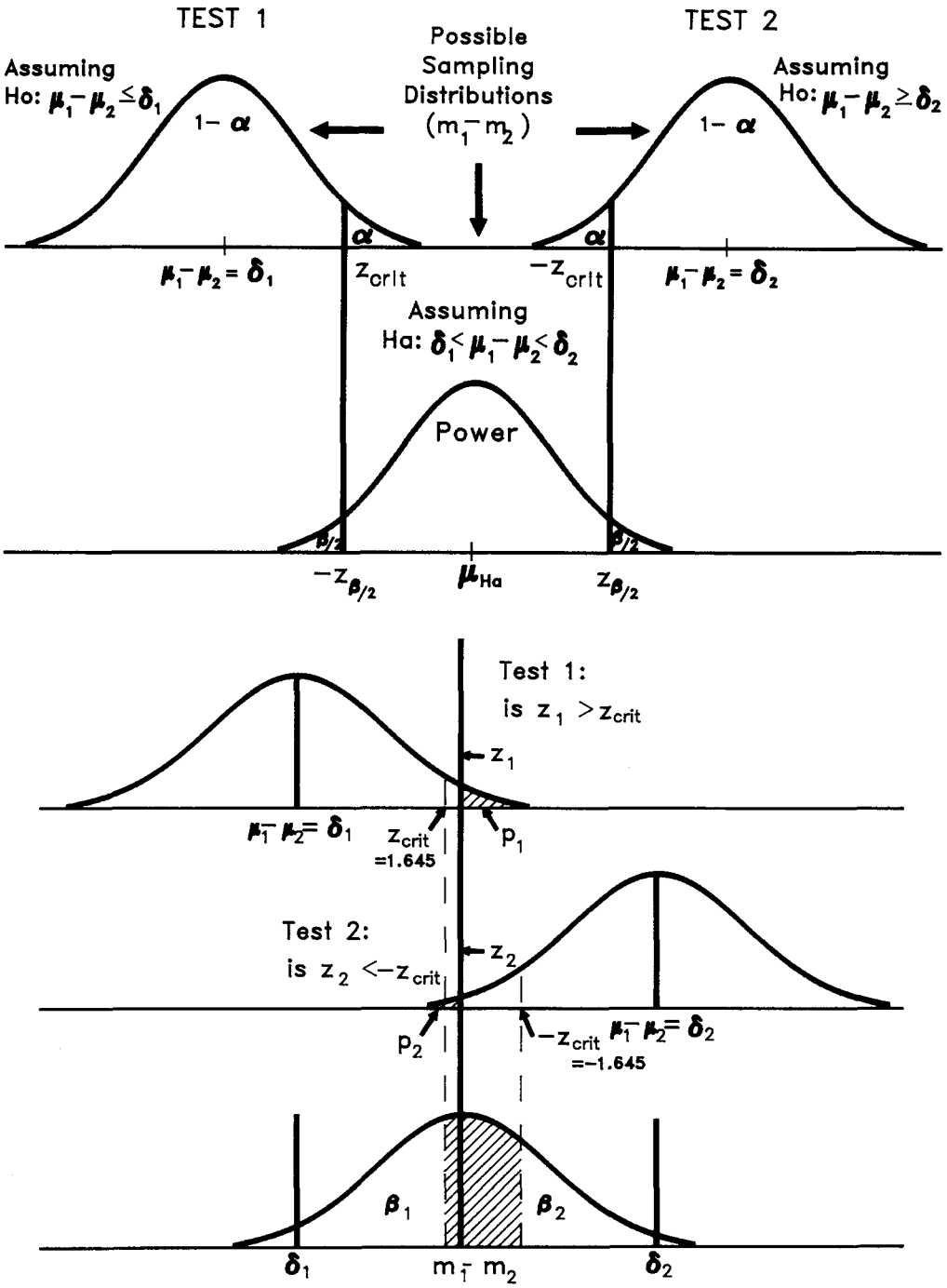

Figure 1. A: Equivalence testing using two one-sided tests. (In this illustration, $\delta_2 = -\delta_1$ and $\mu_{Ha} = 0$. These assumptions are not required [see text]. Ho = null hypothesis; Ha = alternative hypothesis; crit. = critical.) B: Two one-sided tests. (If we assume $M_1 - M_2 = \mu_1 - \mu_2$, then the actual power of this test = $1 - (\beta_1 + \beta_2)$. $z_{crit.}$ = critical z-score value).

confidence interval does not fall within the equivalence interval of ±6.9% (i.e., 20% of 34.3).

To facilitate the interpretation of the results in Table 4, we classified baseline characteristics as different, equivalent, different and equivalent, or equivocal. *Equivocal* status was assigned if neither a statistically significant difference nor a statistically significant equivalence was found. Baseline characteristics were classified as *equivalent* if the two groups were statistically equivalent and not statistically different. Baseline characteristics were classified as *different* if the groups were

Table 1
Hypothesis and Test Statistics to Establish Equivalency Between Two Means or Two Proportions

Parameter	Hypothesis	Test statistic	Rejection criteria
$\mu_1 - \mu_2$	Test 1 $\begin{cases} Ho\text{: } \mu_1 - \mu_2 \le \delta_1 \\ Ha\text{: } \mu_1 - \mu_2 > \delta_1 \end{cases}$	$z_1 = \dfrac{(M_1 - M_2) - \delta_1}{s_{M_1 - M_2}}$	Using significance levels: $p(z_1) \le \alpha$ and $p(z_2) \le \alpha$
	Test 2 $\begin{cases} Ho\text{: } \mu_1 - \mu_2 \ge \delta_2 \\ Ha\text{: } \mu_1 - \mu_2 < \delta_2 \end{cases}$	$z_2 = \dfrac{(M_1 - M_2) - \delta_2}{s_{M_1 - M_2}}$	Using a critical test statistic (z_α): $\lvert z_1 \rvert \ge z_\alpha$ and $\lvert z_2 \rvert \ge z_\alpha$
$p_1 - p_2$	Test 1 $\begin{cases} Ho\text{: } p_1 - p_2 \le \delta_2 \\ Ha\text{: } p_1 - p_2 > \delta_1 \end{cases}$	$z_1 = \dfrac{(\hat{p}_1 - \hat{p}_2) - \delta_1}{s_{\hat{p}_1 - \hat{p}_2}}$	Using a confidence interval: $\delta_1 < [(M_1 - M_2) \pm z_\alpha s_{M_1 - M_2}] < \delta_2$
	Test 2 $\begin{cases} Ho\text{: } p_1 - p_2 \ge \delta_2 \\ Ha\text{: } p_1 - p_2 < \delta_2 \end{cases}$	$z_2 = \dfrac{(\hat{p}_1 - \hat{p}_2) - \delta_2}{s_{\hat{p}_1 - \hat{p}_2}}$	$\delta_1 < [(\hat{p}_1 - \hat{p}_2) \pm z_\alpha s_{\hat{p}_1 - \hat{p}_2}] < \delta_2$

Note.

$$s_{M_1 - M_2} = \left[\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{1/2};$$

$$s_{\hat{p}_1 - \hat{p}_2} = \{[\hat{p}_1(1 - \hat{p}_1)/n_1] + [\hat{p}_2(1 - \hat{p}_2)/n_2]\}^{1/2}.$$

If the parameter of interest is $\mu_1 - \mu_2$, the t distribution should be used if degrees of freedom are small. δ_1 and δ_2 define equivalency. α = probability of a Type I error where for z_α the probability $(z > z_\alpha) \le \alpha$. Alternatively, *Ho* (null hypothesis) and *Ha* (alternative hypothesis) may be expressed as *Ho*: $\mu_1 - \mu_2 \le \delta_1$, or $\mu_1 - \mu_2 \ge \delta_2$: *Ha*: $\delta_1 < \mu_1 - \mu_2 < \delta_2$.

found to be different at a statistically significant level but not statistically equivalent. Finally, a classification of *different and equivalent* was assigned if the groups were determined to be both statistically different and statistically equivalent.

Figure 4 shows the results of this classification. About 48% (13/27 = 48.1%) of the baseline variables were found to be different between the abortion group and the birth group, 37% (10/27 = 37.0%) to be equivocal, 11% (3/27 = 11.1%) to be equivalent, and 4% (1/27 = 3.7%) to be different and equivalent. In the present example, equivalency tests and traditional tests together provide information suggesting that the comparison groups lack baseline similarity.

Note that because a random process cannot be assumed in this quasi-experiment, the p values obtained for both equiva-

Table 2
Traditional and Equivalency Test Results for MMPI Scores of Alcohol Versus Drug-Dependent Subjects

	Alcohol (n = 207)		Drug (n = 49)		Difference			Traditional				Equivalence[b]			
										95% CI				90% CI	
Scale	*M*	*SD*	*M*	*SD*	*M*	*SE*	Equivalence criterion[a]	*z*	*p*	LCL	UCL	*z*	*p*	LCL	UCL
L	49.3	7.1	48.9	9.1	0.4	1.195	±4.93	0.335	.369	−1.941	2.741	−3.792	.000*	−1.565	2.365
F	63.6	8.8	65.2	9.9	−1.6	1.433	±6.36	−1.117	.132	−4.408	1.208	3.322	.000*	−3.957	0.757
K	47.4	6.4	49.5	8.1	−2.1	1.073	±4.74	−1.957	.025†	−4.203	0.003	2.460	.007*	−3.865	−0.335
Hs	66.2	16.6	63.6	15.7	2.6	2.611	±6.62	0.996	.160	−2.517	7.717	−1.540	.062	−1.695	6.895
D	76.7	14.8	68.7	75.1	8.0	2.360	±7.67	3.389	.000†	3.374	12.626	0.140	.556	4.117	11.883
Hy	64.4	12.2	63.1	13.1	1.3	1.966	±6.44	0.661	.254	−2.553	5.153	−2.614	.004*	−1.934	4.534
Pd	70.4	12.3	75.1	12.8	−4.7	1.969	±7.04	−2.387	.009†	−8.560	−0.840	1.188	.117	−7.940	−1.460
Mf	59.2	9.5	61.4	10.9	−2.2	1.554	±5.92	−1.416	.078	−5.245	0.845	2.394	.008*	−4.756	0.356
Pa	59.7	10.5	63.0	10.4	−3.3	1.665	±5.97	−1.982	.024†	−6.564	−0.036	1.603	.054	−6.039	−0.561
Pt	67.5	14.4	67.1	14.9	0.4	2.303	±6.75	0.174	.431	−4.114	4.914	−2.757	.003*	−3.388	4.188
Sc	65.2	16.3	69.7	18.5	−4.5	2.659	±6.52	−1.692	.045	−9.712	0.712	0.760	.224	−8.874	−0.126
Ma	62.5	12.0	70.2	10.2	−7.7	1.856	±6.25	−4.149	.000†	−11.337	−4.063	−0.781	.783	−10.753	−4.647
Si	58.9	10.0	55.4	8.3	3.5	1.541	±5.89	2.271	.012†	0.479	6.521	−1.551	.060	0.965	6.035

Note. Data from Cannon, Bell, Fowler, Penk, and Finkelstein (1990). MMPI = Minnesota Multiphasic Personality Inventory; *L* = Lie; *F* = Frequency; *K* = Correction; *Hs* = Hypochondriasis; *D* = Depression; *Hy* = Hysteria; *Pd* = Psychopathic Deviate; *Mf* = Masculinity–Femininity; *Pa* = Paranoia; *Pt* = Psychasthenia; *Sc* = Schizophrenia; *Ma* = Hypomania; *Si* = Social Introversion; CI = confidence interval; LCL = lower confidence limit; UCL = upper confidence limit.

[a] Criterion is ±10% of the alcohol group mean.

[b] The highest p value of the two one-sided tests has been reported.

* $p < 0.05$ for equivalency, per each one-tailed test. † $p < 0.025$ for traditional test, two-tailed.

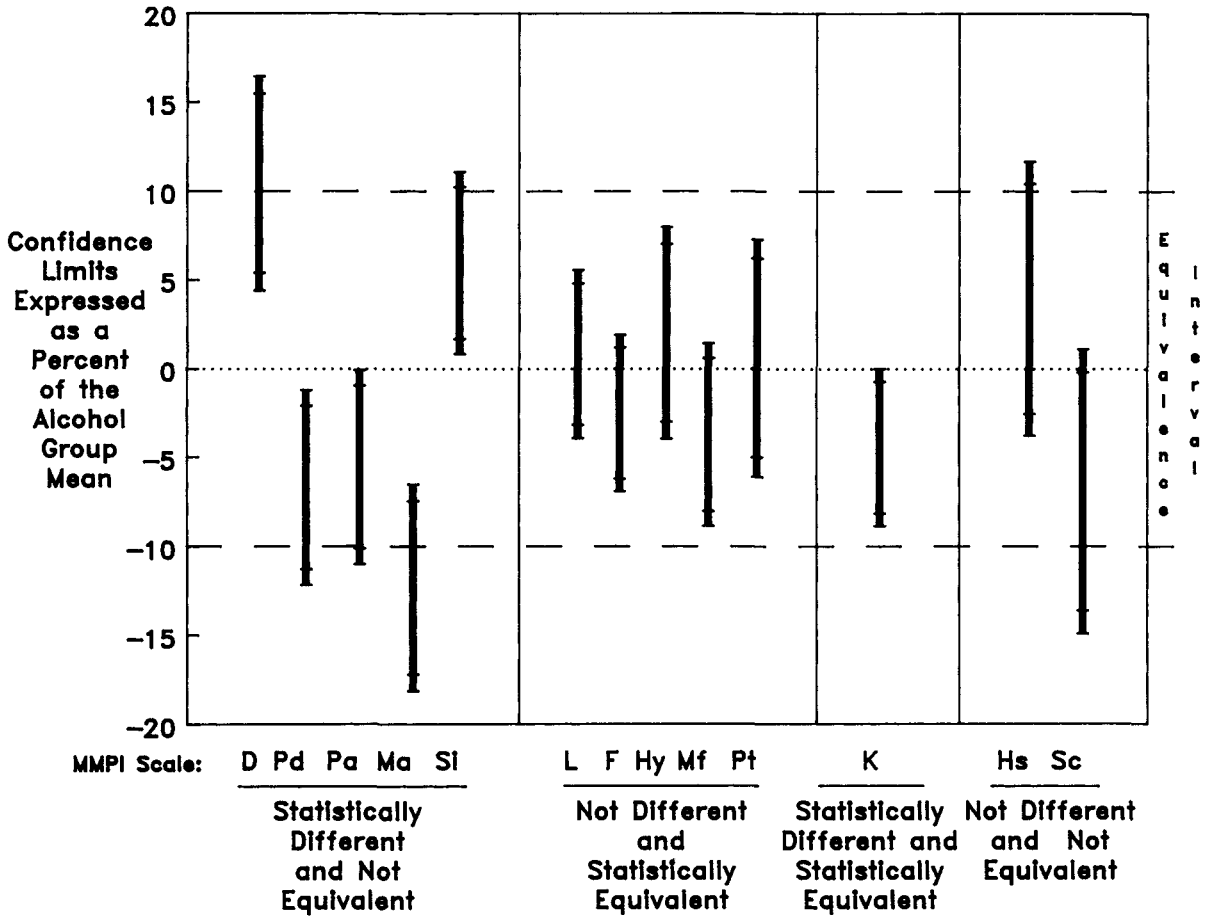

Figure 2. The 90% and 95% confidence intervals around alcohol group means minus drug group means. (Outer tick marks reflect 95% confidence interval. Inner tick marks reflect 90% confidence interval. MMPI = Minnesota Multiphasic Personality Inventory; *D* = Depression; *Pd* = Psychopathic Deviate; *Pa* = Paranoia; *Ma* = Hypomania; *Si* = Social Introversion; *L* = Lie; *F* = Frequency; *Hy* = Hysteria; *Mf* = Masculinity–Femininity; *Pt* = Psychasthenia; *K* = Correction; *Hs* = Hypochondriasis; *Sc* = Schizophrenia).

lency and traditional tests are, technically, descriptive measures of distance, not probabilities. Although *p* values in this example provide a useful measure of similarity (or difference) between baseline measures, *p* values resulting from the equivalency tests do not reflect the probability of a mean difference within an equivalency interval, just as *p* values resulting from the traditional tests do not reflect the probability of a mean difference greater than zero. A probability interpretation would be appropriate only if one were willing to assume that the underlying sampling distribution for the mean difference has resulted from a random process, an unlikely event in the present example.

Discussion

In a traditional test, a mean difference of zero is chosen as the null hypothesis to compute the probability of obtaining the test statistic value. If that probability is sufficiently small, the investigator elects to believe that an incorrect assertion was made and that the population means differ by some amount. However, it is never possible to prove, short of complete enumeration of the populations, that the mean difference is not zero—or to know what it is if it is not zero. From a purely theoretical stance, a small *p* value is expected on occasion, therefore, the fact that a small *p* value exists is of no consequence, one way or the other, to the assertion that the mean difference is or is not zero. As a practical matter, however, if a small *p* value is obtained, the investigator decides that something "unusual" has occurred and the null hypothesis is rejected. However, it is the concomitant occurrence of a small *p* value with a known experimental manipulation after random assignment that changes the experimenter's mind about the null hypothesis, not the *p* value alone.

In an equivalency test, even if the two one-sided tests each result in a small probability that their respective test statistic values have occurred by chance under the assumption that the mean difference in reality is as large or larger than the hypothe-

Table 3
Effect Sizes Comparing Various Therapies in the Treatment of Depression

Scale	Effect size		Traditional				Equivalence[a]			
					95% CI				90% CI	
	M	*SE*	*z*	*p*	LCL	UCL	*z*	*p*	LCL	UCL
Cognitive vs. behavioral	0.12	0.09	1.333	.091	−0.056	0.296	−0.889	.187	−0.028	0.268
Cognitive vs. cognitive–behavioral	−0.03	0.12	−0.250	.401	−0.265	0.205	1.417	.078	−0.227	0.167
Behavioral vs. cognitive–behavioral	−0.16	0.10	−1.600	.055	−0.356	0.036	0.400	.345	−0.325	0.005
Cognitive vs. general verbal	−0.15	0.20	−0.750	.227	−0.542	0.242	0.250	.401	−0.479	0.179
Behavioral vs. general verbal	0.15	0.13	1.154	.124	−0.105	0.405	−0.385	.350	−0.064	0.364
Cognitive–behavioral vs. general verbal	0.09	0.27	0.333	.369	−0.439	0.619	−0.407	.342	−0.354	0.534
Psychotherapy vs. drug therapy	0.07	0.04	1.750	.040	−0.008	0.148	−3.250	.001*	0.004	0.136
Psychotherapy vs. combination	−0.01	0.08	−0.125	.450	−0.167	0.147	2.375	.009*	−0.142	0.122
Combination vs. drug therapy	−0.05	0.21	−0.238	.406	−0.462	0.362	0.714	.238	−0.395	0.295
Psychotherapy vs. tricyclics	0.07	0.04	1.750	.040	−0.008	0.148	−3.250	.001*	0.004	0.136
Psychotherapy vs. combination (tri)	−0.05	0.08	−0.625	.266	−0.207	0.107	1.875	.030*	−0.182	0.082
Combination (tri) vs. tricyclics	−0.05	0.26	−0.192	.424	−0.560	0.460	0.577	.282	−0.478	0.378

Note. Data adapted from Robinson, Berman, and Neimeyer (1990) with further manipulation by James L. Rogers, Kenneth I. Howard, and John T. Vessey. CI = confidence interval; LCL = lower confidence limit; UCL = upper confidence limit; tri = tricyclics.
[a] The equivalency interval uses $\delta = \pm 0.20$; the highest p value of the two one-sided test has been reported.
* $p < 0.05$ for equivalency, per each one-tailed test.

sized value, the investigator cannot, theoretically, conclude that the true difference is within the equivalence interval. As a practical matter, the investigator will elect to believe that the treatments are equivalent when a small p value occurs in the context of a known experimental manipulation after random assignment. In the same way that a traditional test is used within an experimental context to dispel the belief that a difference of zero exists, so an equivalency test is used within an experimental context to rule out the presence of a difference that would make a difference.

If statistical significance has been obtained using a traditional test, the effect size might nevertheless be close enough to zero that, for practical purposes, one decides not to reject the null hypothesis after all but to treat the small difference one believes exists as though it really were zero (i.e., negligible). Again, the use of probability theory in isolation is abandoned. The p value is interpreted in the context of an observed difference between treatment means, that is, the ES. The counterpart to this situation is somewhat changed in equivalency testing because assumed reality under the alternative hypothesis (i.e., the equivalence interval) can arbitrarily be set to be meaningfully large, even when the ES—the distance between an equivalency interval endpoint and the sample mean difference—is small. Said another way, the alternative hypothesis in an equivalency test is that the mean difference falls into a bounded region determined by the investigator, whereas in the traditional test it is not bounded by the investigator but rather constitutes all values except zero. Consequently, practical considerations (as compared with probabilistic considerations) enter at the point of defining a meaningful ES in a traditional test but at the point of defining the equivalence interval in an equivalency test.

The traditional test and the equivalency test are not mutually exclusive. If both tests are conducted, it is possible that both will be rejected, that neither will be rejected, or that one will be rejected and the other will not be rejected. It is instructive to consider the following possibilities.[9]

1. In the event that the equivalency test rejected its null hypothesis, whereas the traditional test failed to reject its null hypothesis, the investigator would conclude that no clinically important difference between the two groups exists.

2. If both null hypotheses were rejected, the investigator would conclude that the treatment difference was larger than the standard null value (usually zero) but smaller than a difference that would make the groups nonequivalent. This outcome has traditionally been addressed through the warning that very large sample sizes ("too much" statistical power) may result in statistical significance even though the ES is clinically trivial (Fleiss, 1981). Equivalency testing provides a more exact method to accomplish this objective. For example, an investigator may be interested in knowing whether a statistically significant but clinically trivial depressive mood shift has occurred. In this case, a change ranging between some nontrivial value and zero might be used as the upper and lower bounds of an equivalence interval in a formal equivalency test.

3. In the event that the traditional test rejected its null hypothesis, whereas the equivalency test failed to reject its null hypothesis, the investigator would conclude that there is a difference between the two groups.

4. It is also possible that both the equivalency and the standard hypothesis tests will fail. In this case, the investigator would conclude that insufficient evidence exists to make any

[9] We are not suggesting that the failure to reject the null hypothesis in either the traditional test or the equivalency test in any way changes one's confidence in a significant result in the other test. This would be falling into the same trap of "proving the null hypothesis" that we are trying to help investigators avoid. We are merely presenting the four possible results an investigator could encounter if both the traditional and equivalency tests were performed on the same data.

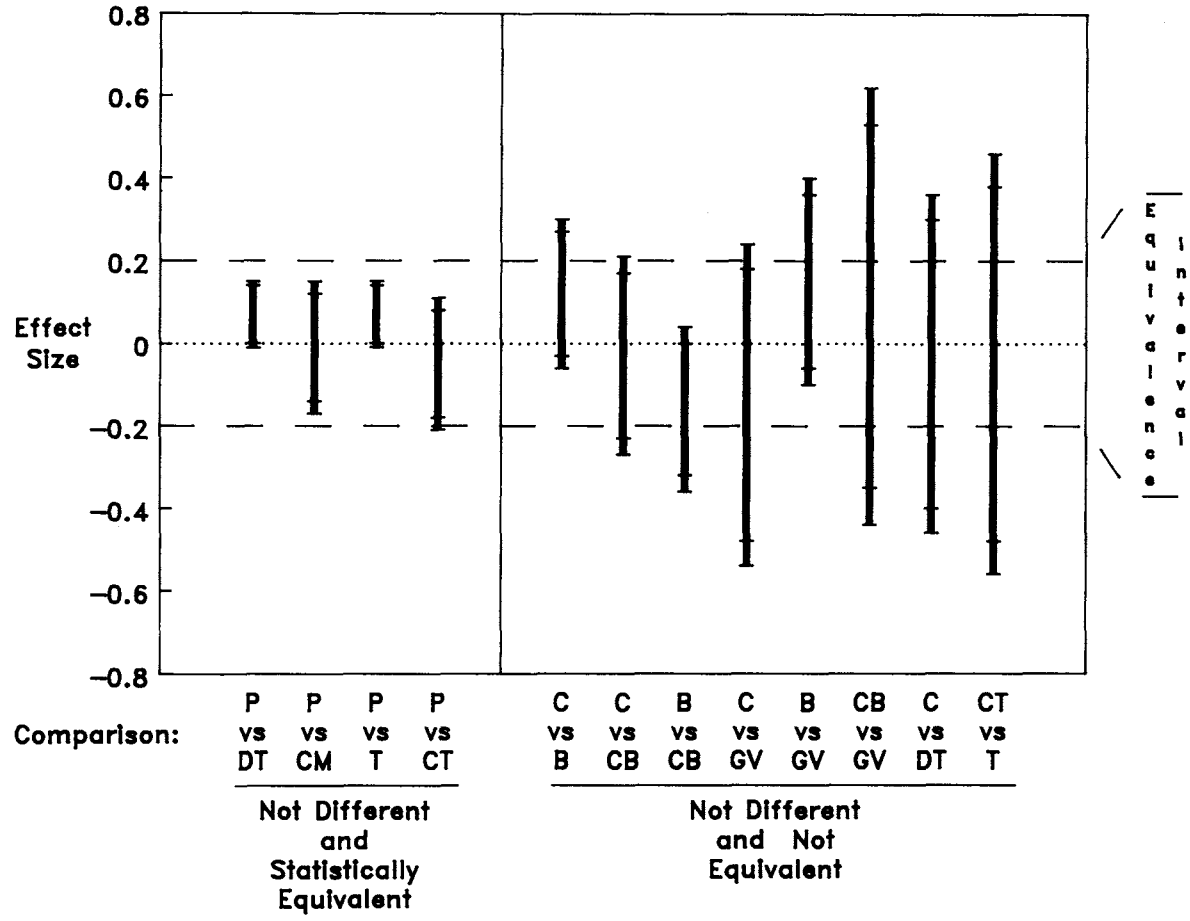

Figure 3. The 90% and 95% confidence intervals around mean effect sizes for various comparisons between therapies used to treat depression. (Outer tick marks reflect 95% confidence interval. Inner tick marks reflect 90% confidence interval. P = Psychotherapy. DT = Drug therapy. CM = Combination. T = Tricyclics. CT = Combination [tricyclics]. C = Cognitive. B = Behavioral. CB = Cognitive–behavioral. GV = General verbal.)

decision. That is, the investigator would surmise that the effect was not reliable enough to conclude either a sizable difference or a reliably small difference.

The latter situation might arise in experiments exhibiting insufficient statistical power because of inadequate sample size or excessive noise (within group variation). For example, an equivalency test failing to indicate a reliably small ES might justify a decision to continue an experimental program that to date had exhibited "negative" findings. Whereas the experimenter might ordinarily conduct a power analysis to facilitate this decision, equivalency testing has two advantages. First, the variance of the estimated difference is used in equivalency testing. In power analysis, the estimated difference is treated as though it was without sampling variation. Second, equivalency testing is conducted using an exact or asymptotic sampling distribution. Power analysis relies on the sample variance as an estimate of the true population.

Controversial issues that surround the relevance and adequacy of the statistical hypothesis test as a means to scientific discovery apply to equivalency tests as well. Some issues that will no doubt arise as investigators consider the role that equivalency testing might play in their research include the following.

1. *Multiple comparisons.* If more than one equivalency test is conducted, the question of adjustment to contain experimentwise alpha will arise. In as much as there are various opinions as to when and how this should be done, the same controversies can be expected to carry over to equivalency testing. In general, the investigator who wishes to control experimentwise alpha should make an appropriate adjustment on the basis of the number of tests actually performed if a priori contrasts are designated, or the number of tests implied on a continuum from least to most likely to be equivalent (i.e., distance between means), if the data are inspected post hoc.

2. *Multiple dependent variables.* Redundant independent measurement should be avoided in research using equivalency tests just as it is avoided in research using traditional tests. When confronted with redundancy in outcome measurements, the investigator should select a priori the most serviceable outcome parameter from each independent outcome dimension or

Table 4
Baseline Differences Between Adolescents Who Elect Abortion and Adolescents Who Carry to Term

Characteristic	Abortion		Birth		Difference		Equivalence criterion[b]	Traditional				Equivalence[a]			
										95% CI				90% CI	
	p	*n*	*p*	*n*	Dif.	*SE*		*z*	*p*	LCL	UCL	*z*	*p*	LCL	UCL
Catholic	.123	141	.043	93	.080	.035	±.025	2.302	.011†	.012	.148	1.594	.945	.023	.137
The "female raiser" was working	.606	141	.495	93	.111	.066	±.121	1.677	.047	−.019	.241	−0.154	.439	.002	.220
Real father present in the home	.135	141	.097	93	.038	.042	±.027	0.903	.183	−.044	.120	0.261	.603	−.031	.107
Curfew, even on the weekend	.723	141	.538	93	.185	.064	±.145	2.892	.002†	.060	.310	0.631	.736	.080	.290
In school last year	.993	141	.978	93	.015	.017	±.199	0.895	.185	−.018	.048	−10.959	.000*	−.013	.043
In school this year	.986	141	.914	93	.072	.031	±.197	2.345	.010†	.012	.132	−4.077	.000*	.021	.123
Ever repeated grade	.343	141	.505	93	−.162	.065	±.069	−2.474	.007†	−.290	−.034	−1.427	.923	−.270	−.054
Think highest grade will be ≤12	.281	141	.500	93	−.219	.064	±.056	−3.411	.000†	−.345	−.093	−2.536	.994	−.325	−.113
Behind grade for age	.355	141	.511	93	−.156	.066	±.071	−2.376	.009†	−.285	−.027	−1.295	.902	−.264	−.048
Used contraception at first intercourse	.511	141	.430	93	.081	.066	±.102	1.220	.111	−.049	.211	−0.319	.375	−.028	.190
Never used contraception	.284	141	.237	93	.047	.058	±.057	0.808	.210	−.067	.161	−0.168	.433	−.049	.143
Used contraception at time of suspected conception	.295	141	.355	93	−.060	.063	±.059	−0.956	.169	−.183	.063	−0.016	.506	−.163	.043
Ever pregnant before	.142	141	.280	93	−.138	.055	±.028	−2.506	.006†	−.246	−.030	−1.990	.977	−.229	−.047
Had a pregnancy test last year	.553	141	.489	93	.064	.067	±.111	0.960	.168	−.067	.195	−0.699	.242	−.046	.174
Wanted a pregnancy	.058	141	.045	93	.013	.029	±.012	0.446	.328	−.044	.070	0.048	.519	−.035	.061
Having a baby would be a problem to me	.922	141	.571	91	.351	.057	±.184	6.203	.000†	.240	.462	2.944	.998	.258	.444
Having an abortion would be a problem to me	.209	139	.557	88	−.348	.063	±.042	−5.507	.000†	−.472	−.224	−4.845	1.000	−.452	−.244
Both having a baby and having an abortion would be a problem to me	.137	139	.256	86	−.119	.055	±.027	−2.149	.016†	−.228	−.010	−1.654	.951	−.210	−.028
Having a baby would be a problem to me, having an abortion would not	.784	139	.314	86	.470	.061	±.157	7.703	.000†	.350	.590	5.133	1.000	.370	.570
Having an abortion would be a problem to me, having a baby would not	.072	139	.291	86	−.219	.054	±.014	−4.081	.000†	−.324	−.114	−3.813	1.000	−.307	−.131
Neither having an abortion or having a baby would be a problem to me	.007	139	.140	86	−.133	.038	±.001	−3.493	.000†	−.208	−.058	−3.456	1.000	−.196	−.070
Abortion is all right if woman was raped	.907	141	.891	92	.016	.041	±.181	0.393	.347	−.064	.096	−4.067	.000*	−.051	.083
Abortion is all right if woman is very young (<15 yr)	.904	141	.750	92	.154	.052	±.181	2.990	.001†	.053	.255	−0.520	.301	.069	.239
Abortion is all right if pregnancy would endanger woman	.838	141	.813	92	.025	.051	±.168	0.489	.312	−.075	.125	−2.788	.003*	−.059	.109
Abortion is all right if woman doesn't want a child now	.788	141	.714	92	.074	.058	±.158	1.268	.102	−.040	.188	−1.433	.076	−.022	.170
Abortion is all right if child is defective or deformed	.620	141	.511	92	.109	.066	±.124	1.646	.050	−.021	.239	−0.226	.410	.000	.218
Abortion is all right if woman wants abortion for any reason	.573	141	.517	92	.056	.067	±.115	0.840	.201	−.075	.187	−0.878	.190	−.054	.166

Note. Data from Zabin, Hirsch, and Boscia (1990). CI = confidence interval; LCL = lower confidence limit; UCL = upper confidence limit; Dif. = difference.
[a] The highest *p* value of the two one-sided tests has been reported. [b] The equivalency interval was defined to be ±20% of the abortion percentage.
* $p \leq .05$ for equivalency, per each one-tailed test. † $p \leq .025$ for traditional test, two-tailed.

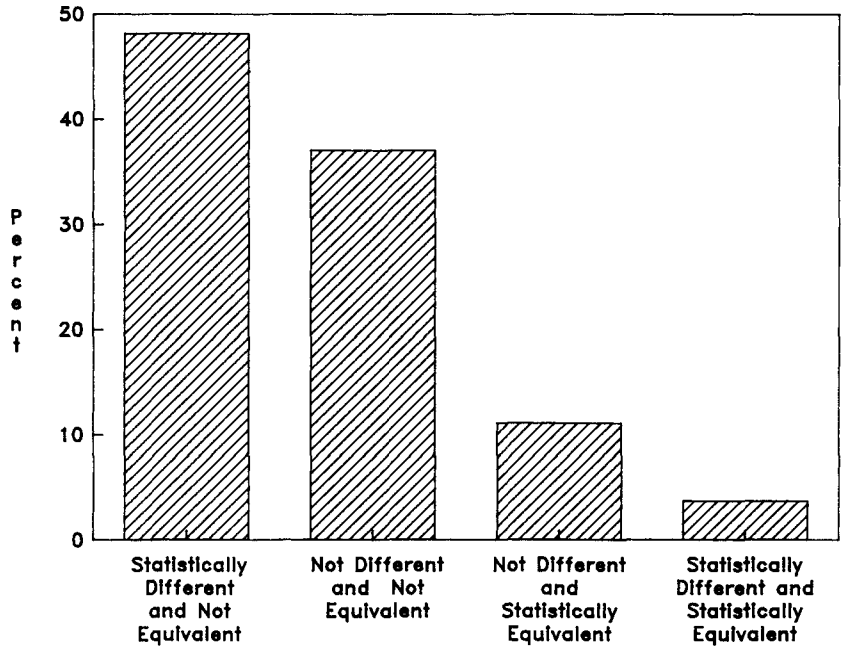

Figure 4. Percentage of baseline characteristic comparisons falling into four possible categories.

collapse the data to a meaningful, relatively independent set of outcome measures using an appropriate preanalysis filter, such as factor analysis or cluster analysis. Extension of the univariate case (presented here) to the multivariate case should be possible and will no doubt constitute a future area of investigation.

3. *Equivalency interval definition.* As with any statistical analysis, equivalency procedures must involve thoughtful planning by the investigator. Meaningful equivalency intervals, sample sizes that reflect appropriate power, and a reasonable Type I error rate will need to be determined. Furthermore, the assumptions underlying the model on which the two one-sided tests or confidence interval are based must be appropriate to the data at hand.

Although these requirements are formidable, they demand no more of the investigator than would be the case if a traditional experimental hypothesis was the goal. In both instances, the investigator will need to determine an appropriate statistical model and address the feasibility of meeting the underlying model assumptions. Furthermore, acceptable Type I error and power levels will have to be decided and a sample size estimated. To determine the sample size requirement for a traditional test, the investigator will need to integrate clinical experience, the work of others, any other available information, and intuition to determine an appropriate ES (i.e., a difference large enough to be important). Analogously for equivalency testing, the investigator, using the same heuristics and reasoning, will need to determine an equivalency interval (i.e., a difference small enough to be unimportant). The thoughtless application of "cookbook" prescriptions is ill-advised, regardless of whether the goal is to establish a difference or to establish equivalency between treatments.

4. *The meaning of* p *values.* Whether generated by a traditional test or an equivalency test, p values will need to be evaluated either within a strict inferential framework or within an exploratory framework. Investigators often use p values as a supplemental measure of distance to aid their post hoc search for relationships among groups and variables. These exploratory activities constitute legitimate scientific undertakings but should always be followed by focused, inferential experimentation to test a well-defined hypothesis. The failure to follow up exploratory procedures with inferential experiments is an indictment against current levels of adequacy in the experimental scientific process but has little bearing on either traditional testing or equivalency testing per se. Both the traditional test and the equivalency test have a role in each of these important aspects of science.

These and other issues will no doubt provide the focus of future work in the area of equivalency testing.

If thoughtfully used, equivalency testing should promote a number of objectives commonly encountered in social scientific research. Some of these objectives are as follows.

1. Substantive issues in clinical and industrial psychology that seek to determine whether an innovation is equivalent to a standard could be formally tested. As it stands, the investigator is often forced, in effect, to accept the null hypothesis if it is not rejected. This is typically justified under the assumption that sufficient statistical power existed to have found a meaningful difference if it were present; that is, the null hypothesis could have been rejected. This approach is convoluted; equivalency testing is, on the other hand, straightforward.

2. Equivalency testing might be used to examine the observed distance between mean baseline values in randomized experiments and quasi-experiments and between mean values for evaluable subjects in randomized experiments evidencing

attrition. However, rejecting a hypothesis of nonequivalence is not a substitute for randomization, a panacea for bias caused by subject attrition, or a justification for rerandomization.

Potential selection bias remains a threat in quasi-experiments and experiments evidencing attrition even when equivalency testing suggests that differences on selected baseline parameters are small. The important point, not altered in the least by the application of either an equivalency or a traditional hypothesis test, remains that quasi-experiments and experiments evidencing attrition typically cannot be presumed to reflect an underlying random process, though both tests require this assumption. In a traditional test, the probability of a large random difference (greater than zero) is obtained, whereas in an equivalency test the probability of a small random difference (less than the equivalency interval) is calculated. Thus, both equivalency and traditional tests are descriptive rather than inferential when applied in these situations.

Similarly, the expected value of the mean difference between samples is zero, regardless of the outcome of equivalency tests performed on baseline parameters in randomized experiments. Thus, rerandomization, when important baseline differences are discovered by traditional tests, equivalency tests, or both, does not eliminate bias but rather introduces it. The failure to establish equivalency may lead the investigator to rerandomize. However, the subsequent sampling distribution would, in fact, be restricted, and the experimenter would be obliged to acknowledge the conservative bias that would result in future traditional tests and the liberal bias that would result in future equivalency tests. Alternatively, the investigator would need to generate an appropriate sampling distribution (mathematically or by computer-intensive methods).

3. Equivalency testing provides a formal procedure to evaluate published negative findings to determine whether these findings also support a reasonable definition of equivalency. As such, equivalency testing might be conducted in place of power analysis, but with the advantages already noted above.

4. Equivalency testing can be used to justify pooling groups. It would replace the informal procedures presently used, such as pooling in the presence of sufficiently high p values (say, .25 or greater).

In effect, equivalency testing provides the investigator with a simple statistical tool that avoids the inappropriate exploitation of nonsignificant results. As such, equivalency testing will advance the scientific process.

References

Adams, W. J. (1974). *The life and times of the central limit theorem.* New York: Kaedmon.

Adler, N. E., David, H. P., Major, B. N., Roth, S. H., Russo, N. F., & Wyatt, G. E. (1990). Psychological responses after abortion. *Science, 248,* 41–44.

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.

Cannon, D. S., Bell, W. E., Fowler, D. R., Penk, W. E., & Finkelstein, A. S. (1990). MMPI differences between alcoholics and drug abusers: Effect of age and race. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 51–55.

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

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Hauck, W. W., & Anderson, S. (1986). A proposal for interpreting and reporting negative studies. *Statistics in Medicine, 5,* 203–209.

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

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108,* 30–49.

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.

Westlake, W. J. (1981). Bioequivalence testing—A need to rethink (Reader reaction response). *Biometrics, 37,* 591–593.

Westlake, W. J. (1988). Bioavailability and bioequivalence of pharmaceutical formulations. In K. E. Peace (Ed.), *Biopharmaceutical statistics for drug development* (pp. 329–352). New York: Marcel Decker.

Zabin, L. S., Hirsch, M. B., & Boscia, J. A. (1990). Differential characteristics of adolescent pregnancy test patients: Abortion, childbearing and negative test groups. *Journal of Adolescent Health Care, 11*(2), 107–113.

Zabin, L. S., Hirsch, M. B., & Emerson, M. R. (1989). When urban adolescents choose abortion: Effects on education, psychological status and subsequent pregnancy. *Family Planning Perspectives, 21,* 248–255.

Received December 11, 1990
Revision received June 16, 1992
Accepted June 19, 1992 ■