Psychological Methods
2006, Vol. 11, No. 4, 386–401

1082-989X/06/$12.00 DOI: 10.1037/1082-989X.11.4.386

When Effect Sizes Disagree: The Case of *r* and *d*

Robert E. McGrath
Fairleigh Dickinson University

Gregory J. Meyer
University of Toledo

The increased use of effect sizes in single studies and meta-analyses raises new questions about statistical inference. Choice of an effect-size index can have a substantial impact on the interpretation of findings. The authors demonstrate the issue by focusing on two popular effect-size measures, the correlation coefficient and the standardized mean difference (e.g., Cohen's *d* or Hedges's *g*), both of which can be used when one variable is dichotomous and the other is quantitative. Although the indices are often practically interchangeable, differences in sensitivity to the base rate or variance of the dichotomous variable can alter conclusions about the magnitude of an effect depending on which statistic is used. Because neither statistic is universally superior, researchers should explicitly consider the importance of base rates to formulate correct inferences and justify the selection of a primary effect-size statistic.

Keywords: effect-size estimation, base rate, correlation coefficient

In recent years, behavioral researchers have witnessed an important change in what is considered optimal statistical practice. With growing awareness of the differences between statistical and practical significance, the importance of power analysis for significance testing, meta-analysis as an integrative strategy, and the limitations of significance testing (e.g., Harlow, Mulaik, & Steiger, 1997; Thompson, 2002), recommendations for incorporating effect-size estimates into statistical analyses have become more definitive. For example, the fourth edition of the *Publication Manual of the American Psychological Association* (American Psychological Association [APA], 1994) "encouraged" authors to report effect sizes in statistical analyses (p. 18). By 1999, Leland Wilkinson and the APA Task Force on Statistical Inference wrote "always present effect sizes for primary outcomes," and "we must stress . . . that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (p. 599). In response to this recommendation, the most recent edition of the APA *Publication Manual* (APA, 2001, p. 25) indicates the reporting of effect sizes is "almost always necessary." More than 20 journals in the field of behavioral research now require authors to report effect-size statistics, at least for key statistical analyses (a list is provided at http://www.coe.tamu.edu/~bthompson). It is no longer considered sufficient to ask of an effect or relationship: "Is it there?" It is increasingly considered essential to also ask "How much is there?" and sometimes even "Is it enough to care?"

The purpose of this article is to examine an important but often overlooked complication to the use of effect sizes. Depending on the character of the variables under investigation, the researcher may have several reasonable choices of effect-size measures. However, because alternative statistics describe the relationship between two variables in different ways, they can lead to different conclusions about the size or importance of that relationship, even when one effect-size measure can be directly converted into the other. This represents a significant obstacle to the objective interpretation of research findings.

To illustrate the problem, we focus on a comparison of two popular effect-size measures used to examine relationships between a dichotomous and a quantitative variable, Cohen's *d* and the point-biserial correlation coefficient (r_{pb}). They were chosen for several reasons. First, they are familiar to most researchers and are among the most commonly used effect-size measures in meta-analytic investigations in the behavioral sciences. Though new effect-size measures continue to be introduced, many if not most research studies continue to rely on these well-known statistics. Second, unlike many effect-size statistics, benchmarks for interpreting the size of these effects have been proposed (Cohen, 1988) and widely adopted. Though there are some problems with these benchmarks, discussed below, their existence enhances the illustration of the issues involved. Third, r_{pb} and *d* are mathematically appropriate to the same universe of analytic situations. Fourth, formulas are readily available for converting one measure to the other, which fosters the impression they are interchangeable, even though they may lead to very different conclusions about the same data.

Robert E. McGrath, School of Psychology, Fairleigh Dickinson University; Gregory J. Meyer, Department of Psychology, University of Toledo.

Portions of this article were presented at the 2003 midwinter meeting of the Society for Personality Assessment, San Francisco, CA. Both authors contributed equally to this article.

Correspondence concerning this article should be addressed to Robert E. McGrath, School of Psychology T-WH1-01, Fairleigh Dickinson University, Teaneck Hackensack Campus, Teaneck, NJ 07666. E-mail: mcgrath@fdu.edu

Computational Issues

Cohen's *d* conceptualizes relationships between dichotomous and quantitative variables in terms of the difference between the quantitative variable means for the two groups defined by the dichotomous variable. This difference is standardized by using the within-group standard deviation pooled across the two groups. Cohen's *d* is particularly popular in experimental and quasi-experimental research in which a difference in the effect of a treatment or manipulation on group means is considered important. The standardized mean difference actually encompasses a set of related statistics, the presentation of which is unfortunately complicated by discrepancies in the symbols used by different authors. The lack of standardization in the literature may reflect the relative recency of interest in these statistics. In an effort to clarify this and future discussions of standardized mean differences, Table 1 summarizes symbols used in five key discussions of these statistics and our recommendations for a standard.[1]

The standardized mean difference in the population is computed from the following formula:

$$\delta = \frac{\mu_{\cdot 1} - \mu_{\cdot 2}}{\sigma_{\cdot j}}, \tag{1}$$

where $\mu_{\cdot 1}$ and $\mu_{\cdot 2}$ designate the means for the two populations represented in the sample by the two groups, whereas $\sigma_{\cdot j}$ is the standard deviation for either population under the assumption of equal variances. The corresponding formula for the sample standardized mean difference is

$$d = \frac{\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2}}{S_{pooled}}, \tag{2}$$

where $\overline{Y}_{\cdot 1}$ and $\overline{Y}_{\cdot 2}$ designate the means for the two groups. S_{pooled} refers to the standard deviation generated by summing together the sums of squares for the two sample groups, dividing by *N* (the total sample size), and taking the square root of the resulting variance. Hedges and Olkin (1985) noted this formula for *d* also provides the maximum likelihood estimate of δ for a sample.

In practice, sample estimation of δ is usually based on the least squares estimator:

$$g = \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{\hat{\sigma}_{pooled}} = d\sqrt{\frac{N-2}{N}}. \tag{3}$$

$\hat{\sigma}_{pooled}$ involves dividing the pooled sums of squares by $N - 2$ instead of by *N*. Dividing the pooled sums of squares by $N - 2$ corrects for bias in the sample estimate of $\sigma_{\cdot j}$, whereas dividing it by *N* is consistent with the definitional formula of a standard deviation. $\hat{\sigma}$ is probably the most common quantity used for purposes of inference in general, and the meta-analysis of group differences in particular. This is because effect sizes are often computed by using variances or standard deviations generated by statistical software such as SPSS and SAS, which by default calculate $\hat{\sigma}$, with $N - 1$ in the denominator, rather than *S*, with *N* in the denominator. Hedges and Olkin (1985) demonstrated that *g* is a biased estimator of δ. The best sample estimate of δ can be approximated by using the following formula (as simplified by Hunter & Schmidt, 2004):

$$\hat{\delta} = \frac{g(N-3)}{N-2.25} = d\left(\frac{N-3}{N-2.25}\right)\sqrt{\frac{N-2}{N}}. \tag{4}$$

The existence of three sample statistics for computing the standardized mean difference can lead to confusion, especially given discrepancies in the use of labels. In the following, we focus primarily on *d*, as defined in Equation 2, for several reasons. The three sample statistics differ only slightly and converge as sample size increases. The relatively small practical difference between formulas may also explain why the sample correlation coefficient, which is also a biased statistic (Fisher, 1915), is rarely corrected in practice.[2] In addition, the use of the descriptive rather than inferential versions of both *d* and r_{pb} simplifies the formulas used to relate the standardized mean difference to r_{pb}. At the same time, the issues illustrated in this article that use *d* apply equally to the whole family of standardized mean differences presented in Table 1.

The correlation coefficient is a particularly popular effect size in observational studies or individual differences research in which the goal is to estimate the validity of one variable as a predictor of the other or the magnitude of association between two variables. Because common symbols for the population (ρ) and sample (*r*) correlations are generally accepted, the discussion moves directly to the point-biserial correlation, which is simply the standard Pearson product–moment correlation coefficient applied to the case of a dichotomous and a quantitative variable. The

[1] There are other versions of standardized mean difference statistics as well. The statistic developed by Glass (1976), which divides the mean difference by the standard deviation of the control group, is excluded as it is generally considered inferior to those considered here (e.g., Hunter & Schmidt, 2004). We only include formulas that assume equality of variances in the two populations, as these are the most commonly used in practice. With unequal variances, the conversions provided between r_{pb} and *d* are not exact.

[2] The unbiased estimate of the population correlation, ρ, is given by the adjusted *r*,

$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(N-1)}{N-2}},$$

where *N* indicates the total sample size. This equation is the simplified version of the more familiar R_{adj} formula used in multiple regression,

$$R_{adj} = \sqrt{1 - \frac{(1-R^2)(N-1)}{N-k-1}},$$

where *k* is the number of variables predicting the criterion variable.

Table 1
Comparison of Symbols Used for Standardized Mean Difference Statistics

Study	Parameter	Sample statistics: Pooled within sample sums of squares divided by: N	Sample statistics: Pooled within sample sums of squares divided by: $N - 2$	Corrected for bias
Cohen (1988)	d		d_s	
Hedges & Olkin (1985)	δ	$\hat{\delta}$[a]	g	d
Hunter & Schmidt (2004)	δ	d[a]	d	d*
Lipsey & Wilson (2001)			ES_{SM}	
Rosenthal (1991)	d	d	g	g^u
Recommended	δ	d	g	$\hat{\delta}$

[a] Discussed only as the maximum likelihood estimate of the population value.

r_{pb} conceptualizes relationships in terms of the degree to which variability in the quantitative variable and the dichotomous variable overlap.

One standard formula for the point-biserial correlation as a descriptive rather than inferential statistic is as follows:

$$r_{pb} = \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{S_Y}\sqrt{p_1 p_2}. \quad (5)$$

S_Y is the standard deviation generated by dividing the total sums of squares for the quantitative variable by N. When $\overline{Y}_{\cdot 1} \neq \overline{Y}_{\cdot 2}$, S_Y is larger than S_{pooled}, the standard deviation used to compute d (Equation 2), and the size of the difference between the two standard deviations is directly related to the size of the difference between the means (demonstrated in the Appendix). As a result, the correlation is bounded within the interval −1.00 to 1.00.[3] The formula also includes the terms p_1 and p_2, which indicate the base rates or proportions of participants in each of the dichotomous variable groups, with $p_2 = 1 - p_1$.

The Effect of Base-Rate Inequalities

The reason d and r_{pb} can lead to different conclusions can be demonstrated several different ways. As the difference between p_1 and p_2 in Equation 5 increases, their product becomes smaller, so r_{pb} decreases. Because p_1 and p_2 are not part of the formula for d, the latter statistic is unaffected by base-rate disparities. As a result, d and r_{pb} differ markedly in terms of the degree to which they are affected by the base rate for the two values of the dichotomous variable. Thus, r_{pb} can be understood as a base-rate-sensitive effect-size measure, whereas d is base-rate-insensitive.

This difference in sensitivity to base rates can also be stated in terms of the variance of the dichotomous variable. Because the variance of this variable is a function of the product of the base rates (i.e., with the dichotomous groups coded as two consecutive numbers such as 0 and 1, $S_X^2 = p_1 p_2$ and $S_X = \sqrt{p_1 p_2}$), variance is maximized when $p_1 = p_2 = .50$. As the proportions become more discrepant, the variance of the dichotomous variable becomes smaller (see Figure 1, left vertical axis), resulting in a decline in the value of the correlation similar to that resulting from range restriction. Thus, rather than saying r_{pb} is base-rate-sensitive and d is base-rate-insensitive, one could just as readily state that r_{pb} is a variance-sensitive effect-size measure, whereas d is variance-insensitive. In this case, it is important to remember that the variance referred to is that of the dichotomous variable not the within-group or total variance for the quantitative variable. Goodman (1991) also suggested the terms marginal-dependent and marginal-free to represent the two classes of statistics.

In pursuit of making the difference between d and r_{pb} even clearer, the standard formulas can be modified to illustrate the two critical distinctions:

$$d = \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{\sqrt{S^2_{pooled}}} \quad (6)$$

and

[3] The true possible range of the point-biserial correlation is actually smaller. No correlation can reach a value of 1.00 unless the two variables have the same distribution. Because quantitative and dichotomous variables by definition have different distributions, the true range for the point-biserial correlation is always less than 1.00 to −1.00 and varies depending on the distribution of the quantitative variable. For example, Nunnally and Bernstein (1994) reported the point-biserial correlation is restricted to the interval from −.79 to .79 if the quantitative variable is normally distributed and continuous.

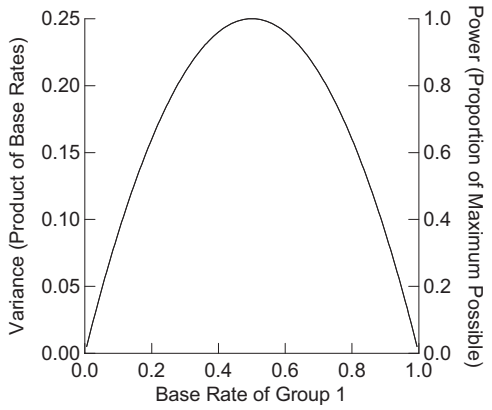

Figure 1. The left vertical axis indicates the variance of the dichotomous variable as a function of $p_1\,p_2$. This quantity reaches its maximum (.25) when $p_1 = p_2 = .50$. The right axis indicates the relative power of d as a function of $p_1\,p_2$. These values are four times those in the left axis.

$$r_{pb} = \frac{(\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2})}{\sqrt{\frac{S^2_{pooled}}{p_1 p_2} + (\bar{Y}_{\cdot 1} - \bar{Y}_{\cdot 2})^2}}. \quad (7)$$

Equation 7 may be unfamiliar, so its derivation is provided in the Appendix. Now the denominator in both cases is based on S^2_{pooled}. The inclusion of the differences between the means in the denominator of Equation 7 serves to place r_{pb} on a scale that ranges from -1.00 to 1.00, in contrast with d, which is unbounded. The equation also highlights the less commonly known distinction, which is that the pooled variance is divided by p_1p_2 in the computation of r_{pb}. As the difference between p_1 and p_2 increases, the value by which S^2_{pooled} is divided becomes smaller, as does r_{pb} when compared with d.

Still a third approach to understanding the latter difference between d and r_{pb} is available by using the standard formula for converting from d to the correlation coefficient (Lipsey & Wilson, 2001),[4]

$$r_{pb} = \frac{d}{\sqrt{d^2 + \frac{1}{p_1 p_2}}}. \quad (8)$$

As in Equation 7, the denominator repeats the numerator term to restrict values to the range -1.00 to 1.00. Inclusion of the inverse of the product of the base rates means that for any value of d, the corresponding r_{pb} gets smaller as the base rate becomes more extreme. In contrast, the formula for converting from r_{pb} to d is as follows:

$$d = \frac{r_{pb}}{\sqrt{(1 - r^2_{pb})\,p_1 p_2}}. \quad (9)$$

In this case, the product of the base rates produces a smaller denominator term as the rates become more disparate, which produces a corresponding increase in d.

What are the practical implications of the differential sensitivity that r_{pb} and d have to the base rate or variance of the dichotomous variable? If schizophrenia only occurs in 1% of the population, a base-rate-sensitive effect size takes into account the heightened difficulty of differentiating or predicting this rare event, such that the magnitude of the effect is much smaller than it would be if instead schizophrenia occurred in 50% of the population. As a result, to obtain an accurate base-rate-sensitive effect size that will generalize to the population, researchers must ensure that sample base rates mirror the population base rates. In contrast, a base-rate-insensitive effect size treats the base rate (or variance) of the dichotomous variable as irrelevant. In the case of schizophrenia, a base-rate-insensitive effect size would be of similar magnitude regardless of whether the disorder affected 1% or 50% of the population and, thus, regardless of whether it was generated from a sample in which 1% or 50% of the participants met criteria for the diagnosis.

For fixed values of p_1 and p_2, the choice between r_{pb} and d is arbitrary with respect to the rank ordering of effect-size values. If the d value relating dichotomous variable X to quantitative variable Y is larger than the d value relating X to quantitative variable Z, then the correlation between X and Y must also exceed the correlation between X and Z. To state this principle more generally, if there are two pairs of dichotomous and quantitative variables, and the base rates for the dichotomous variables are the same, then the rank ordering of the two correlations will be the same as the rank ordering of d values. However, this relationship dissolves once the dichotomous variables differ in terms of their base rates.[5]

Table 2 demonstrates both these points. For a given value of p_1 and p_2, a larger d value is consistently associated with

[4] This equation offers a good example of some of the confusion that has resulted from inconsistency in the use of symbols associated with the standardized mean difference statistics. Aaron, Kromrey, and Ferron (1998) considered the formula provided in the text for computing r from d to be inaccurate and offered the following alternative:

$$r_{pb} = \frac{d}{\sqrt{d^2 + \frac{N^2 - 2N}{n_1 n_2}}},$$

where n_1 and n_2 refer to the group sizes. However, the Aaron et al. formula is actually the correct formula for converting g to the correlation coefficient. Once the use of symbols is clarified, the discrepancy is resolved. The formula here converts g to r_{pb}; equation 8 in the text converts d to r_{pb}.

[5] Kraemer et al. (1999) demonstrated this principle holds for other effect-size statistics as well.

Table 2
Conversions Between d and r_{pb} as a Function of Differences in Base Rates or Variance

Base-rate disparity	Variance of the dichotomous variable	p_1	p_2	d	r_{pb}
		Large d			
Large	Small	.98	.02	0.80	.11
Moderate	Moderate	.75	.25	0.80	.33
None	Maximum	.50	.50	0.80	.37
		Medium d			
Large	Small	.98	.02	0.50	.07
Moderate	Moderate	.75	.25	0.50	.21
None	Maximum	.50	.50	0.50	.24
		Small d			
Large	Small	.98	.02	0.20	.03
Moderate	Moderate	.75	.25	0.20	.09
None	Maximum	.50	.50	0.20	.10

a larger r_{pb} value. For example, when $p_1 = .75$, a d of 0.80 is associated with r_{pb} of .33, whereas a d of 0.50 is associated with r_{pb} of .21. Thus, when p_1 is constant, the rank ordering of effect sizes is preserved across the two measures.

This relationship no longer exists when p_1 varies across analyses. For instance, consider the first three rows in Table 2 when $p_1 = .98$, $d = 0.80$ is associated with $r_{pb} = .11$. However, as p_1 approaches .50, r_{pb} increases so that $r_{pb} = .33$ when $p_1 = .75$, and $r_{pb} = .37$ when $p_1 = .50$. Even though d did not change, r_{pb} increased when the difference in base rates was less extreme. Furthermore, what is often considered a large d value (i.e., 0.80; Cohen, 1988) is associated with a small value for r (i.e., .11), when the probability of one of the two dichotomous values is only .02. A base rate of .02 (2 cases per 100) may seem like an extremely rare outcome, but in fact it is not. For instance, many psychiatric conditions have a prevalence of .02 or less in the general population, including dysthymia, agoraphobia, panic disorder, bipolar disorder, schizophrenia, any drug use disorder, or any specific personality disorder (Narrow, Rae, Robins, & Regier, 2002; Torgersen, Kringlen, & Cramer, 2001). The same is true for numerous medical conditions. For example, a recent study found that 2.3% of older males with normal levels of prostate-specific antigen had a serious form of prostate cancer upon biopsy (Thompson et al., 2004). It is also likely that many social and experimental phenomena commonly studied by psychologists are similarly infrequent, though—as we discuss below—it is often difficult to estimate the true frequency of these events.

Table 2 demonstrates another impediment to achieving comparable results with r_{pb} and d, though not for mathematical reasons. Many users of Cohen's (1988) benchmarks seem unaware that those for the correlation coefficient and d are not strictly equivalent, because Cohen's generally cited benchmarks for the correlation were intended for the infrequently used biserial correlation rather than for the point biserial. This creates a slight advantage for d over r in terms of the characterization of effect sizes when those benchmarks are used for other types of correlation coefficients. For example, as demonstrated in the table, when base rates are equal, the d value Cohen suggested as large (0.80) corresponds to an r_{pb} value of .37, far less than his commonly cited benchmark for a large r value (.50). To achieve comparability between r_{pb} and d when base rates are equal, the benchmarks for small, medium, and large correlations would need to be changed to .10, .24, and .37, respectively (Cohen, 1988, pp. 22, 82; Lipsey & Wilson, 2001, p. 147). Alternatively, to equate d with r_{pb}, the benchmarks for d would need to be changed to 0.20, 0.67, and 1.15, respectively. The former would seem the better option, as surveys of the empirical literature suggest that the Cohen benchmarks are in fact too high for correlation coefficients in general, considering the magnitude of effects commonly found in research (Hemphill, 2003; Richard, Bond, & Stokes-Zoota, 2003).

Base rate or variance sensitivity is a feature of many other statistics besides correlations. In fact, all hypothesis testing statistics (e.g., t, χ^2, F) are base-rate-sensitive. This means that base-rate-sensitive effect-size statistics more accurately track the power of hypothesis tests than do base-rate-insensitive effect sizes. For example, as the difference between p_1 and p_2 increases, the value of the independent groups t test—and therefore its power—declines for a given mean difference. The point-biserial correlation also declines, but d does not. The following equation for t helps to illustrate to this relationship (equivalent to Rosenthal, 1991, Equation 2.6),

$$t_{(df)} = \frac{(\bar{Y}_{.1} - \bar{Y}_{.2})}{S_{pooled}} \sqrt{(df)\, p_1 p_2}, \tag{10}$$

where the new term, df, indicates the degrees of freedom ($N - 2$). This equation can be considered in light of both the structurally similar Equation 2 for computing d and Equation 5 for computing r_{pb}. Equation 10 differs from both the preceding equations in the inclusion of the degrees of freedom, in that t values increase as degrees of freedom increase for a given mean difference. The only other difference from the r_{pb} formula is the inclusion of S_{pooled} in the denominator rather than S_Y. As noted previously, the latter bounds r in the range -1.00 to $+1.00$; t is not similarly bounded.

With respect to d, the only other difference from Equation 2 is the standard deviation of the dichotomous variable. As with r_{pb}, t is reduced to the extent that p_1 and p_2 are discrepant for a given mean difference. As the base rates for the dichotomous variable become more disparate or, equivalently, as the variance in the dichotomous variable becomes more restricted, t declines even though the magnitude of the standardized mean difference is unchanged.

Though we are restricting our discussion of the impact of base rates to the case of d and r_{pb}, it is noteworthy that the base-rate issue has broader implications. The negative effect of disparate base rates on the probability of correct inferences in clinical settings (e.g., Dawes, 1962; Meehl & Rosen, 1955) and correct classifications in selection (e.g., Rorer, Hoffman, LaForge, & Hsieh, 1966; Schmidt, 1974; Taylor & Russell, 1939) has been known and discussed for years. However, the effect of base rates on effect sizes is probably less familiar. Among effect-size measures appropriate to cases in which both variables are dichotomous, the odds and risk ratios are relatively insensitive to the base rate of the predictor variable, as are sensitivity and specificity, whereas the correlation (phi) coefficient, chi-square, absolute risk reduction, and the number needed to treat are base rate sensitive (for descriptions of these statistics, see Barratt et al., 2004; Streiner, 2003).

Other statistics are base rate sensitive, but the impact of their sensitivity varies. Positive predictive power (PPP) is a diagnostic efficiency statistic used in circumstances where both a predictor and outcome variable are dichotomous (Streiner, 2003). When a diagnostic screen is conducted for lung cancer, the PPP of the test is the probability that a person has cancer if the test finding is positive. PPP is affected by the degree to which base rates are unequal in the outcome variable: The lower the probability of cancer in the population of interest, the lower the PPP. If the test were used in a population in which the majority had cancer, PPP would be greater. The opposite is true for negative predictive power (NPP), which is the probability that an individual does not have cancer if the test is negative (Streiner, 2003). NPP is higher when cancer is rare and lower when it is common.

The kappa coefficient—a reliability statistic often used when both variables are dichotomous—is also base-rate- or variance-sensitive, though the impact of a discrepancy in the marginal distributions is more idiosyncratic (Streiner, 2003). If two raters consider the same of two options to be the less likely, the value of kappa will on average be reduced, with the reduction increasing as the base rates of the ratings become more unequal. If both raters use one option infrequently, but differ in terms of which they consider the infrequent option, the value of kappa instead increases on average (Zwick, 1988). Furthermore, unless the sample size is very large, kappa can become extremely unstable if the base-rate inequality is severe, varying between very low and very high values across samples.

Meta-Analytic Examples of Inconsistent Inferences From r and d

Real-world research in meta-analysis can be used to demonstrate the impact of unequal base rates on the interpretation of findings. Christensen, Hadzi-Pavlovic, and Jacomb (1991) reported an average effect size of $d = 1.87$ for the use of neuropsychological tests to differentiate patients with dementia from normal controls. The authors did not report the base rate of dementia in their primary studies, but this would translate into $r_{pb} = .68$ if the dementia base rate was .50. In contrast, if one assumed a more likely dementia base rate of .10, the average correlation would drop to $r_{pb} = .49$. In a screening setting where the base rate of dementia matches that in the general population over age 65 (3%), the average validity of neuropsychological tests to differentiate patients with dementia from normal controls would drop to $r_{pb} = .30$ (Meyer et al., 1998).

Table 3 provides examples of effect-size estimates that were summarized in a previous article (see Meyer et al., 2001). Each entry provides the results of a meta-analysis or a series of studies that used one dichotomous variable and one variable treated as quantitative. They were chosen for use here because the original reference allowed r and d values to be computed for each study and because the studies afforded a spectrum of base rates.

The mean effect sizes were reported as r values by Meyer et al. (2001). To illustrate the impact of base-rate or variance sensitivity, we also computed d. In addition, we used standard formulas (Rosenthal, 1991; Rosenthal, Rosnow, & Rubin, 2000) to generate estimates of what r would equal if base rates were not a factor and what d would equal if they were. A version of r unaffected by the base rates was generated from d by acting as if p_1 were equal to p_2. An artificial version of d that is affected by base rates was also computed. Instead of using Equation 9, which would correctly convert r to d, we computed a d value directly from the actual base-rate-sensitive r values without correcting for discrepancies in the base rates (see the note to the table for more details). Table 3 also indicates the percentage difference in effect-size magnitude when the impact of base rate on r is eliminated, and when base rates influence the value for d.

The findings in Table 3 are ordered by the base rate of the targeted condition, which makes the impact of base-rate sensitivity readily apparent. For example, Entry 1 indicates that only 1 in 100 psychiatric patients ultimately committed suicide. Under these conditions, r would increase by 342% (more than quadruple) if it were computed like d, without regard to the actual frequency of suicides or as if half the patients committed suicide. Conversely, the value of d would decrease by 79% if d were as sensitive as r to the infrequency of suicide in the population.

In contrast, when the base rate of a target condition (or its complement) represents about 30–50% of the population under study, the differences between r and d are not particularly notable. For instance, the mean base rate of cognitive impairment across samples was .37 in Forster and Leckliter's (1994) meta-analysis that examined the ability of the Halstead–Reitan neuropsychological battery to detect cognitive impairment in children (Table 3, Entry 4). In this case, base-rate-sensitive and insensitive effect sizes differed

Table 3
Examples From Meyer et al. (2001) of the Impact of Base-Rate Sensitivity on Effect Sizes

Study description	Study base rate	Effect-size estimates: *r* Actual value	*r* Estimate if base-rate insensitive	*r* % change	*d* Actual value	*d* Estimate if base-rate sensitive	*d* % change
1. Suicide predicted by Beck Hopelessness Scale scores in psychiatric patients (Beck et al., 1985, 1990).	.013	.077	.341	+341.7	0.727	0.156	−78.6
2. Development of breast cancer predicted by denial/repressive coping style (McKenna et al., 1999).	.077	.033	.075	+129.3	0.151	0.067	−55.7
3. Suicide attempts and serotonin metabolites in cerebrospinal fluid in psychiatric patients (Lester, 1995).[a]	.171	.217	.265	+21.8	0.771	0.496	−35.8
4. Differentiating impaired versus control children with Halstead–Reitan Neuropsychological Tests (Forster & Leckliter, 1994).[b]	.370	.321	.353	+9.9	0.800	0.707	−11.6
5. Detecting malingered psychopathology with MMPI validity scales (primarily from analog studies, Berry et al., 1991; Rogers et al., 1994).[b,c]	.542	.741	.746	+0.7	2.675	2.633	−1.6
6. Detecting underreported psychopathology with MMPI validity scales (primarily from analog studies, Baer et al., 1992).[b]	.621	.387	.389	+0.4	0.944	0.940	−0.5

Note. Criterion variables were all dichotomous. Base-rate and effect-size values are averages weighted by sample size. Transformations from one effect size to the other were computed on the original study findings not on the values presented in the table. The estimate if base-rate-insensitive *r* values were computed from the actual *d* values by using the formula $r = d/(d^2 + 4)^{1/2}$. This equation falsely assumes the base rates were equal. The estimate if base-rate-sensitive *d* values were computed from the actual *r* values by using the formula $d = 2r/(1 - r^2)^{1/2}$. % change was computed from nine-digit summary effect-size values and not the three-digit values reported in the table. MMPI = Minnesota Multiphasic Personality Inventory.
[a] An error with one effect size in the original meta-analysis was corrected. [b] The base rate indicated is not a natural base rate; these studies overselected one of the two groups of interest. [c] *r* was computed from *d*. In the other studies, both *r* and *d* were computed directly from descriptive or inferential statistics.

by only about 10%. However, it is important to recognize that for purposes of enhancing power, researchers often oversample target cases or use equal-sized target and control groups, which may seriously underestimate the degree of base-rate inequality in the population. For instance, it is questionable whether 37% represents a reasonable estimate of how frequently cognitive impairment occurs in many applied settings. If testing was being conducted in an educational setting in which just 10% of the children were expected to have some form of cognitive impairment, the validity coefficient should drop from $r = .32$ to $r = .23$.

A similar analysis can be applied to the experimental study of psychological phenomena, though the comparison is often complicated by the lack of information about the true base rates for the events studied. To illustrate, we provide an example from social psychology. Carlson, Marcus-Newhall, and Miller (1990) presented a meta-analysis of studies investigating whether aggressive cues facilitate aggressive responding in negatively toned situations. They found that aggressive responding was greater when a weapon was present than when it was not, as long as there was no evidence the participants were aware of the research hypothesis; the mean *d* value was 0.31. Most of the studies they cited used equally sized groups, even those Anderson, Lindsay, and Bushman (1999) later classified as field studies that should generalize to everyday life. Because these experiments were intended to provide insight into actual social phenomena, it is reasonable to ask how well the presence of a weapon predicts the intensity of aggression in the real world, a question that is usually addressed using *r* rather than *d*. However, to do so would require an estimate of the base rate for a weapon cue in society at large. It would be difficult to do so precisely, though the base rate is undoubtedly less than .50. Figure 2 indicates how the mean correlation would decline as the base rate of a weapon cue decreases from the .50 value used in these experiments to almost 0. The estimated correlations suggest that the degree to which a weapon cue accounts for the intensity of aggressive acts, and therefore its importance as a real-world predictor of aggression, varies as a function of its frequency in the population. This cannot be estimated accurately from experimental designs with artificially equated base rates or from base-rate-insensitive effect sizes. However, as will be discussed later, effectiveness as a predictor is still not the same thing as importance as a risk factor.

Is Base-Rate Sensitivity Good or Bad?

Some have argued that base-rate sensitivity reduces the utility of a statistic. For example, several commentators have objected to the kappa coefficient specifically because its base-rate sensitivity often makes the reliability of ratings

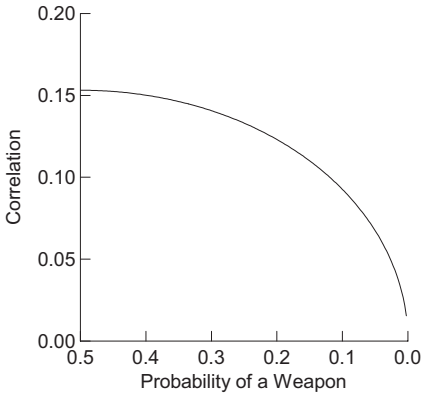

Figure 2. The average *d* for studies evaluating the facilitation of aggression by the presence of a weapon among potentially unaware participants was 0.31. The corresponding average *r* decreases from a maximum of about .15 to near zero as a function of decreasing the base rate for the presence of a weapon cue in the natural environment.

look poor (Brennan & Prediger, 1981; Spitznagel & Helzer, 1985; Zwick, 1988). Others have contended in response that base-rate sensitivity is appropriate to a reliability statistic (Bartko, 1991; Shrout, Spitzer, & Fleiss, 1987). As the base-rate inequality increases, total variance decreases. This means that measurement error variance will tend to increase as a proportion of the total variance, and reliability in fact decreases.

Haddock, Rindskopf, and Shadish (1998) argued for the odds ratio on the basis of its insensitivity to the distribution of dichotomous variables. Some researchers have conducted comparisons of effect-size measures under the assumption that effect sizes should not be affected by variable distributions (Hunter, 1973; von Eye & Mun, 2003); others have not made this assumption (Costner, 1965; Kraemer et al., 1999). It is not surprising that the former have criticized the correlation coefficient, whereas the latter have been more supportive of its use. For example, Kraemer et al. (1999) were troubled by the odds ratio's indirect relationship to power, a criticism that as noted above can be leveled at *d* and at all other base-rate-insensitive effect-size measures.[6] Complicating matters is the possibility, to be discussed below, that the purposes of the analysis may be an important factor in deciding between base-rate-sensitive and insensitive statistics.

Unfortunately, standard texts that behavioral researchers rely upon for guidance in the use of effect sizes often provide little or no information about this controversy. Cohen's (1988) discussion of power analysis noted the practical impact that base rates have on the use of power tables for *d* as well as the difference in the effect of base rates on *d* and *r*, but he offered no guidance about choosing between the two options. Lipsey and Wilson's (2001) guide to meta-analysis recommends *d* unequivocally when the dichotomous variable base rates are markedly different. Rosenthal (1991) did not address the issue formally, but he clearly considers *r* the more useful of the two because of its greater flexibility for characterizing associations across diverse research designs and statistical analyses. However, even though Rosenthal et al.'s (2000) guide to calculating effect sizes for linear contrasts emphasizes *r*, it also includes many formulas that compute a base-rate-insensitive version of *r*. Hunter and Schmidt (2004) recommended adjusting the point-biserial correlation when base rates are unequal, though these corrections are not commonly used in practice.[7]

Comparing the Merits of *r* and *d*

As the preceding discussion suggests, the choice between a base-rate-sensitive effect size like r_{pb} and a base-rate-insensitive one like *d* is not necessarily clear-cut. Below we summarize arguments supporting the superiority of each.

Advantages of r Over d

1. The relationship between *r* and power is more direct.

Though widely criticized, significance testing remains a critical component of inferential practice in psychological research, especially for purposes of interpreting results of an isolated study. Effect sizes are important not only as estimates of the strength of an effect or relationship but also as a component of power analysis (Cohen, 1988). Given that all significance testing statistics are base-rate-sensitive, for a given *N* there is a direct relationship between *r* and significance test results. This is not true for *d* (Rosenthal, 1991), where the relationship to power is moderated by the disparity in group sizes and not just by the total sample size. This difference suggests that with sample size held constant, r_{pb}

[6] Though the statistic has its advocates (e.g., Haddock et al., 1998; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003), others have been quite critical of various aspects of the odds ratio (Davies, Crombie, & Tavakoli, 1998; Kraemer et al., 1999; Sackett, Deeks, & Altman, 1996; von Eye & Mun, 2003). This probably should be considered a controversial statistic at the present time.

[7] Because of r_{pb}'s sensitivity to base rates, the biserial correlation has been suggested as a superior statistic to the point-biserial correlation (e.g., Carroll, 1961), and biserial or polyserial correlations are recommended or even automatically generated for use in structural equation modeling. However, not only does the biserial correlation equate base rates in a way that may not generalize to a real-world population, this statistic requires introducing the assumption of a normally distributed variable underlying the dichotomization, which does not apply to truly categorical variables. Furthermore, inspection of the biserial correlation formula (Cohen et al., 2003) demonstrates the statistic is still somewhat base–rate-sensitive.

is a better indicator of the p value resulting from the corresponding significance test.

It also suggested to us that standard discussions of the relationship between d and power have been remiss in overlooking the issue of base rates. Most textbooks acknowledge the impact of total sample size and alpha level on power. We know of none that includes extreme base rates in a dichotomous variable in the list of moderators of the relationship between effect size and power for base-rate-insensitive statistics, even though it often might be more important in practice than alpha level. Rosnow, Rosenthal, & Rubin (2000; Equation 10) provide an index of the relationship between base-rate inequality and subsequent loss of power, loss $= 1 - (n_h/\bar{n})$, where n_h is the harmonic mean of the group sizes and $\bar{n}$ is their arithmetic mean. When converted to base-rate notation, the formula indicates that the relative loss of statistical power from unequal sample sizes is $1 - 4p_1 p_2$. Thus, when base rates are equal, there is no loss of power [$1 - 4(.5)(.5) = 0$]. However, when 95% of the participants are in one group and 5% are in the other, power declines by 81% [$1 - 4(.95)(.05) = .81$]. Because variance in the dichotomous variable is determined by $p_1 p_2$, it also can be seen that, all other factors being equal, power is directly proportional to the variances indicated in Figure 1. That is, power is at a maximum (and relative loss is at a minimum) in the center of the figure when p_1 and $p_2 = .50$ but drops precipitously as the base rates diverge. Thus, Figure 1 simultaneously illustrates constraints on the size of r_{pb} (left vertical axis) and the relative power associated with d (right vertical axis). Discussions of power should straightforwardly indicate the role of base rates in power analysis. In the absence of this information, the uninformed user can easily overestimate the power of the study based on d.

2. r is a more flexible statistic.

The correlation coefficient can be computed for any combination of dichotomous and quantitative variables. This is an extremely useful characteristic when attempting to make comparisons across a variety of study designs, as is sometimes the case in meta-analysis. Rosenthal (1991) and colleagues (Rosenthal et al., 2000) have provided methods for the use of r as a general indicator of magnitude that is applicable in almost any study. This sort of flexibility is unusual among statistics. Though d can be used when both variables are dichotomous (Haddock et al., 1998), it cannot be used when both are quantitative.

The greater practical flexibility of r corresponds to the broader relevance of the concept of association or relationship versus group differences. In almost any circumstance in which a researcher is interested in considering two variables in conjunction with one another, that interest can be conceptualized in terms of an association between the variables. In contrast, the concept of differences between groups represents a special case of understanding contingent relationships.

3. r is integral to general linear models.

r is a cornerstone of multiple regression statistics, including the standard error of estimate, the regression coefficient, and the index of association. Indeed, r is central to the general linear model in all its forms. This relationship makes r a much more useful statistic than d when the goal of the analysis is prediction of an outcome (Costner, 1965).

One implication of this relationship is that even a dichotomous variable associated with a large d value may not be a particularly useful predictor when the base rates are very different. For example, Entry 6 of Table 3 suggests that individuals who are instructed to underreport pathology on the Minnesota Multiphasic Personality Inventory (MMPI) produce validity indicator scores that are on average about one standard deviation higher ($d = 0.94$) than those generated under normal instructions, a large effect. Suppose that in voluntary psychiatric settings only 2% of respondents have a vested interest in appearing overly healthy. Faking then turns out to have little relationship to respondents' scores ($r = .131$), even with the same standardized mean difference. Consequently, in the absence of information about the true base rate of underreporting, covarying or removing potentially invalid cases could result in unacceptably small improvements in scale validity (e.g., Piedmont, McCrae, Riemann, & Angleitner, 2000).

4. r is dependent on base rates, which has interpretive meaning in applied settings.

It has been suggested that the effect of base rate on the correlation coefficient can have interpretive value.

> Constraints on correlations associated with differences in distribution inherent in the constructs are not artifacts but have real interpretive meaning. . . . The observed correlation between smoking and lung cancer is about .10 . . . . There is no artifact of distribution here; even though the risk of cancer is about 11 times as high for smokers, the vast majority of both smokers and nonsmokers alike will not contract lung cancer, and the relationship is low because of the nonassociation in these many cases. (Cohen, Cohen, West, & Aiken, 2003, p. 54)

Similarly, as the proportion of individuals attempting to underreport on the MMPI declines, the proportion of false positive cases increases. r_{pb} changes to reflect this decline in predictive power. Consistent with the literature on maximizing the accuracy of diagnostic inferences (e.g., Meehl & Rosen, 1955), this makes r a more ecologically valid indicator of the effectiveness of the dichotomous variable as a predictor of the outcome than d when the true base rate is considered. This is a particularly valuable feature of the

correlation coefficient as an indicator of the extent to which one variable can achieve practical utility as a predictor of another.

Even so, it would be a mistake to use the correlation coefficient as sufficient evidence of the relative importance of a risk factor. For example, more than half the American population is now considered overweight if not obese (Flegal, Carroll, Ogden, & Johnson, 2002). If the proportion of overweight adults continues to rise (diverging more from a base rate of .50), the correlation between being overweight and medical complications associated with excess weight in the general population will actually decline, even as weight continues to increase in importance as a risk factor.

Advantages of d Over r

1. Mean differences are particularly relevant for experimental or treatment effects.

Just as the nature of *r* makes it a more useful statistic when the goal is to determine the relationship between a predictor and a criterion, the nature of *d* makes it a more useful and readily understood statistic when the goal is simply to determine the amount of difference in the impact of two experimental conditions or treatments. However, as was noted in connection with the discussion of Figure 2, *d* is not the best indicator of the overall societal impact of an intervention if the population of individuals who receive the treatment is small relative to the total population.

2. *d* behaves more intuitively.

The sensitivity to base-rate differences can lead to some counterintuitive results for *r*. For example, suppose after preliminary analysis a researcher decides to increase the sample size as a means of increasing power. If the subsequent recruitment rate varies across groups and exacerbates a difference in base rates, the overall correlation can actually decline as a result of recruiting, though *d* does not. On the other hand, a decline in *r* can be used to warn the researcher that the sampling method is inefficient.

A second case of base-rate sensitivity producing unexpected results can occur when subgroups are combined. In a recent study (Blanchard, McGrath, Pogge, & Khadivi, 2003), college students completed the MMPI under instructions either to "fake bad" in a manner appropriate to mimic the results for someone not guilty by reason of insanity (forensic feigners) or to achieve psychiatric hospitalization (psychiatric feigners). These groups were then compared with psychiatric patients who completed the MMPI under standard instructions (see Table 4). When forensic feigners were compared with psychiatric patients on eight indicators of malingering, the mean *d* value was 1.98, whereas the mean *d* for comparing psychiatric feigners to psychiatric patients was 2.39. When both groups of feigners were combined in a composite analysis, the mean *d* value was 2.20, falling between the two subgroup means as one would expect.

Across the same eight predictors, the mean correlation between group membership and scale score was .39 for forensic feigners and .49 for psychiatric feigners. However, when the two groups were combined, so that the number of feigners was doubled, the base rate of feigners increased from .053 in the forensic condition and .061 in the psychiatric condition to .107 in the combined condition. Rather than falling between the two correlations based on subgroups of feigners, because the differences in the base rates of the dichotomous variable had declined (and the variance

Table 4
An Instance When Combining Experimental Groups Has a Counterintuitive Impact on r_{pb}

MMPI indicator of "faking bad"[a]	*d*			r_{pb}		
	Forensic feigners (n = 24)	Psychiatric feigners (n = 28)	All feigners (n = 52)	Forensic feigners (n = 24)	Psychiatric feigners (n = 28)	All feigners (n = 52)
F	2.07	2.35	2.24	.42	.49	.57
Fb	1.41	1.71	1.62	.30	.38	.45
Ds	2.19	2.54	2.42	.44	.52	.60
Ds–R	1.78	2.22	2.01	.37	.47	.53
Fp	3.10	3.21	3.16	.57	.61	.70
FBS	0.59	1.41	1.01	.13	.32	.30
O–S	1.35	2.04	1.72	.29	.44	.47
F–K	3.35	3.67	3.44	.60	.66	.73
M	1.98	2.39	2.20	.39	.49	.54

Note. *N* represents the number of feigners in the analysis. In all cases, the comparison group consisted of 432 psychiatric inpatients, who completed the inventory under standard instructions. MMPI = Minnesota Multiphasic Personality Inventory. F = Infrequency; Fb = F Back; Ds = Dissimulation; Ds–R = Dissimulation—Revised; Fp = Frequency-Psychopathology; FBS = Fake Bad; O–S = Obvious–Subtle; F–K = F minus K.
[a] Blanchard et al. (2003).

had increased), the mean correlation increased substantially to .54.[8]

3. *d* estimates effects independent of base rates.

A case may be made for base-rate-insensitive statistics as a general indicator of effect size when the base rate is subject to change across time and situation. Suppose the goal is to estimate the degree to which psychotherapy has been helpful for depression. If *r* is used to evaluate the relationship between treatment choice and ratings of improvement, the statistic will lose generalizability as the proportion of the population of depressives who have received treatment changes. In addition, to the extent that base rates fluctuate from sample to sample for nonsubstantive reasons when conducting a meta-analysis, one would expect greater confounding variability across studies in *r* (which responds to these nonsubstantive fluctuations), when compared with *d* (which does not).

As a result, *d* can provide a better estimate of the "transportability" of an effect to an alternative context where the base rates differ. For instance, parental susceptibility to stress may have a very small association, as measured by *r*, with the incidence of child physical abuse when studied in the general population where the incidence of abuse is quite low. These findings would suggest that interventions designed to bolster coping and stress resistance in parents may have little practical value for actually reducing abuse. However, if the same finding is accompanied by a relatively large *d* value, it would suggest that parental susceptibility to stress is nonetheless relatively important in the limited number of cases in which abuse actually occurs. As such, the *d* value accurately reveals that the stress–abuse relationship will become more apparent in settings in which the base rate for abuse is higher, suggesting, for example, that parental susceptibility to stress should be a more meaningful target of intervention for families in many clinical or forensic settings. As noted previously, the lack of sensitivity to base-rate change has by itself led some writers to prefer base-rate-insensitive statistics.

Choosing What to Report and How to Interpret the Effects

So both statistics have some desirable characteristics. How then is one to proceed? Some of the discussion suggests *r* is particularly suited for cases in which the task is to evaluate criterion-related validity. *d* is more appropriate when the goal is to determine the effect of an intervention or experimental manipulation. Furthermore, still other statistics may be more appropriate when the issue has to do with risk factors for negative outcomes. At times the distinction between these contexts may not be straightforward though. For example, though most of the studies that have evaluated the effectiveness of the MMPI as an indicator of faking good or faking bad have used experimental designs, these are analog studies of a prediction problem, and so *r* would typically be the more appropriate effect-size indicator assuming an ecologically valid estimate of the base rate is available. Similarly, even in experimental social research, in which *d* is the more commonly used effect size, the ultimate goal can be the prediction of real-world outcomes (e.g., Anderson et al., 1999; Funder & Ozer, 1983), a goal for which *r* is again defensibly the better measure. The preceding discussion leads us to the following recommendations, which apply both to individual studies and to meta-analytic summaries.

First, in studies examining causal effects relating a dichotomous variable to a quantitative one, the *d* statistic provides meaningful information. However, if any value is to be gained from evaluating the causal variable as a predictor of the outcome in the real world, and the base rates for the two values of the causal variable are unequal in real life, then the point-biserial correlation can provide distinctly meaningful information as well and so should be computed and reported by using appropriate target base rates.

Second, it is not unusual for studies to equalize the base rates for the dichotomous variable, even when they are very different in the real world, a strategy that distorts the information provided by r_{pb}. This problem is easily addressed if there is a reasonable estimate available of the true base rate in the population simply through the use of the population instead of the sample base rate to generate r_{pb}. For instance, if a sample base rate is .50 but a reasonable estimate of the population base rate is .10, one could easily compute a more accurate estimate of the population correlation coefficient with Equation 5, 7, or 8 by using .10 as p_1.

Third, base rate considerations raise several issues concerning appropriate benchmarking for interpreting effect sizes. As noted above, the commonly cited benchmarks for *r* were intended for use with the biserial correlation (Cohen, 1988) and are too conservative for the correlation coefficient in general. A more reasonable strategy would treat .10 as a small effect, .24 as a moderate effect, and .37 as a large one.

As the base rates for the dichotomous variable become more unequal, the point-biserial correlation and standardized mean respond very differently, and the issues surrounding interpretive benchmarking become more complicated. The correlation coefficient becomes smaller, whereas *d* is unaffected. An effect that the standardized mean difference suggests is substantial can in fact prove to have a trivial

[8] Though we focus on cases of one dichotomous variable, *r* also has the unfortunate tendency in studies of two dichotomous variables (when *r* is typically called the ϕ coefficient) of changing depending on whether the frequencies for one variable are artificially equalized and which variable is selected for equalization (Dawes, 1993; Fleiss, 1981).

impact in real-world situations. A predictor that the correlation coefficient suggests is fairly weak can in fact prove to be quite powerful when considered in light of the inherent difficulty of predicting a rare phenomenon.

Several different approaches to the interpretation of the effect size can be suggested that take these multiple perspectives into account. For example, Rosenthal and Rubin (1982) recommended the binomial effect-size display as a general indicator of the true size of an effect regardless of the distributions of any dichotomous variables involved. However, this argument has been strongly criticized (e.g., Hsu, 2004).

Instead of relying on newer statistics, two options are available that use the familiar d and r_{pb} statistics. One option would involve the use of the standard fixed interpretive benchmarks for small, medium, and large effects suggested by Cohen (1988) as well as complementary interpretive benchmarks that are adjusted in consideration of base rates. Specifically, adjusted interpretive benchmarks for r_{pb} can be obtained by inserting base-rate information and the standard d interpretive benchmark values into the following formula:

$$r_{\textit{Base-Rate-Adjusted Interpretive Benchmark}} = \frac{d_{\textit{Standard Interpretive Benchmark}}}{\sqrt{d^2_{\textit{Standard Interpretive Benchmark}} + \frac{1}{p_1 p_2}}}. \quad (11)$$

Table 5 provides adjustments of the benchmarks for several base rates.[9] So the last three entries in the row labeled *Large* indicate the r values that correspond to $d = .80$ when the base rate increasingly departs from .50 as indicated. To provide an example of the use of the adjusted benchmarks, Table 3 provides information concerning the prediction of suicide by using a Hopelessness scale. The point-biserial correlation was only .077, which is a small effect according to standard benchmarks. However, because the proportion of study participants who actually committed suicide was only .013, the standardized mean difference was 0.727, which is a large effect according to standard benchmarks. Similarly, by using Equation 11, the adjusted interpretive benchmarks for small, medium, and large correlations when one of the base rates equals .013 become .023, .057, and .090, respectively. When considered in light of standard benchmarks for desirable levels of predictive accuracy, the scale is not very effective at predicting suicide. This is because the majority of individuals with high scores on the hopelessness scale do not commit suicide. However, within the limits of predictability created by the extremely disparate real-world base rates, this is nonetheless also a fairly effective predictor. Relative to other measures attempting to predict phenomena of such infrequency, the adjusted interpretive benchmarks suggest that this Hopelessness scale is likely to prove relatively useful despite an observed r of .077.

Both of these interpretive statements reveal something important about the characterization of this effect; they provide complementary perspectives for understanding the relationship between predictor and criterion. The benchmarks are used to characterize the effect represented by the correlation coefficient not the correlation coefficient itself. Clearly, in an absolute sense, a correlation of .077 is small. Even so, this predictor is likely to be of relative value (assuming the criterion is important to predict) given the limits of predictability in this particular context.

Following a similar approach, if base rates are actually deemed important, adjusted interpretive benchmarks for d that consider the base rate can be computed by inserting base-rate information and the standard r_{pb} interpretive benchmark values into the following formula:

$$d_{\textit{Base-Rate-Adjusted Interpretive Benchmark}} = \frac{r_{\textit{Standard Interpretive Benchmark}}}{\sqrt{(1 - r^2_{\textit{Standard Interpretive Benchmark}})\, p_1 p_2}}. \quad (12)$$

Table 6 provides a sample of the adjusted benchmarks for several base rates. For the *Large* row, the last three entries

Table 5
A Sample of Interpretive Benchmarks for r_{pb} Adjusted for Base Rates

Interpretive standard	Standard values		Adjusted benchmarks when $p_1 =$			
	d	r	.50	.75	.95	.99
Large	0.80	.37	.37	.33	.17	.08
Medium	0.50	.24	.24	.21	.11	.05
Small	0.20	.10	.10	.09	.04	.02

Note. Adjusted benchmarks for r_{pb} are derived from the equation $r_{\textit{Base-Rate-Adjusted Interpretive Benchmark}} = \frac{d_{\textit{Standard Interpretive Benchmark}}}{\sqrt{d^2_{\textit{Standard Interpretive Benchmark}} + \frac{1}{p_1 p_2}}}$ and suggest, for example, that if the base rate is .99, even a correlation of .08 indicates a relatively substantial predictive relationship.

[9] Given the links we have already demonstrated between r and t, a parallel that may be helpful to some readers—though we are not suggesting this as an inferential strategy—would be to modify the alpha benchmark required for the statistical significance of t, such that at more extreme base rates, findings would be considered significant at a higher alpha level (e.g., $p = .25$ rather than $p = .05$). This kind of alpha modification parallels what is done by Equation 11 and in Table 5. Conversely, given the links we have demonstrated between d and t, it may help to note that what is done by Equation 12 and illustrated in Table 6 is analogous to imposing onto d the standard alpha level requirement of $p = .05$ for the significance of t. That is, for a fixed total N to achieve a statistically significant t value, an increasingly large mean difference is required as p_1 becomes smaller, and this is the impact that Equation 12 has on d.

Table 6
A Sample of Interpretive Benchmarks for d Adjusted for Base Rates

Interpretive standard	Standard values		Adjusted benchmarks when p_1 =			
	d	*r*	.50	.75	.95	.99
Large	0.80	.37	0.80	0.92	1.84	4.02
Medium	0.50	.24	0.50	0.58	1.15	2.51
Small	0.20	.10	0.20	0.23	0.46	1.01

Note. Adjusted benchmarks for *d* are derived from the equation $d_{Base\text{-}Rate\text{-}Adjusted\ Interpretive\ Benchmark} = \frac{r_{Standard\ Interpretive\ Benchmark}}{\sqrt{(1 - r^2_{Standard\ Interpretive\ Benchmark})p_1p_2}}$ and suggest, for example, that if the base rate is .99, a standardized mean difference of 2.51 is necessary before a predictive relationship can be considered medium sized.

indicate the *d* values that correspond to $r = .37$ when the base rate increasingly departs from .50 as indicated. For instance, the table shows that if the benchmarks for *d* are adjusted in light of base rates, for a treatment to produce what is considered a large standardized mean difference when only 1% of the target population members receive the treatment and the researcher deems that base-rate information is important to the interpretation of the effect, the magnitude of *d* would need to be about 4.0 rather than 0.80.

A final option is to report *d* as well as *r*. Doing so has several benefits, including simplicity and the fact that it does not require adjusting interpretive benchmarks. An additional benefit is that when base rates diverge, reporting both *r* and *d* will juxtapose the seemingly discrepant inferences about magnitude of effect and will highlight the importance of deciding whether the natural base rates should be given credence or be discounted. However, for efficiency, researchers may prefer adjusting the base rates in instances in which large numbers of effect-size statistics are reported for a single sample.

Implications for Meta-Analyses

Meta-analytic researchers have the choice of summarizing research findings with either *r* or *d* whenever the research question involves one dichotomous and one quantitative variable. They also face a similar choice when summarizing findings from the 2×2 data matrices formed by two dichotomous variables. In these instances, researchers need to choose whether to use a base-rate-sensitive statistic such as phi (i.e., *r*), number needed to treat, PPP, or the absolute risk reduction over a base-rate-insensitive statistic such as *d*, the odds ratio, sensitivity, or the relative risk reduction. Rather than relying on what is traditional practice in an area of research, the choice should be made deliberately after carefully considering the issues outlined above. The critical question is one of accurate generalization: Is base-rate sensitivity important for accurately modeling the impact of a predictor, risk factor, intervention, or treatment in the real-life setting where the finding will be applied? If it is, the meta-analyst faces the additional burden of considering the base rate as part of the effect-size estimation. For instance, if one wishes to predict relatively rare real-world outcomes (e.g., malingering, recidivism, diagnosis, employee theft, success in a highly selective training program), each study effect size should be computed by using a reasonable estimate of real-life base rates. Because primary studies vary in the extent to which the sample base rate matches the intended population base rate, it is incumbent on the meta-analyst to select a target base rate and compute effect sizes accordingly (e.g., computing r_{pb} from Equations 5, 7, or 8 by using the targeted base rate rather than the sample base rate).

At the same time, however, as noted in passing several times above, the realistic base rate itself may be a moving target. For instance, a predictor can be used both in a general screening setting (as a risk factor for a disorder among people in the general population) and also in one or more alternative settings in which the condition to be predicted has a higher relative base rate (e.g., patients being screened in a primary care setting; patients being admitted to a tertiary-care hospital specializing in the disorder). Under these circumstances, it would be optimal if the meta-analyst were able to estimate the base rate for each common setting and to provide relevant effect size estimates for each.

Conclusions

Effect sizes have in recent years come into wider use, as meta-analysis has become the integrative strategy of choice among behavioral researchers. Important decisions about clinical and theoretical questions are being made regularly on the basis of these statistics. To date, these investigations have proceeded with little consideration of the impact the statistic of choice has on the outcomes.

Although we welcome the emerging model of inferential judgment based on effect sizes as well as significance tests, we caution the consumers of effect-size statistics that their use is not always as straightforward as it may seem from text descriptions or from traditional practice in any area of research. Depending on whether the goal of an analysis is to estimate the relative size of the impact of one variable on another, the effectiveness of one variable as a real-world predictor of another, or the importance of one variable as a risk factor for another, we would argue that the optimal approach to understanding the effect can vary. There will even be circumstances when it is interesting and important to consider a relationship from more than one of these perspectives.

We have suggested these multiple perspectives on the interpretation of an effect can be achieved through the use

of both *r* and *d* or through simultaneous comparison of one statistic to both standard and adjusted interpretive benchmarks. With regard to the latter possibility, we are reminded of the concerns Cohen (1988) raised with the introduction of his benchmarks.

> The terms "small," "medium," and "large" are relative not only to each other but to the area of behavioral science or even more particularly to the specific content and research method being employed . . . [T]here is a certain risk inherent in offering conventional operational definitions for these terms for use . . . in as diverse a field of inquiry as behavioral science. (p. 25)

Although some progress has been made in suggesting benchmarks that are appropriate to specific areas of behavioral investigation (e.g., Hemphill, 2003; Richard et al., 2003), the preceding discussion suggests that base rates also can be used to adjust benchmarks to the situation. At the same time, it is not our intention to suggest that all effects based on disparate base rates should be interpreted from multiple perspectives. The researcher should evaluate whether it is important to understand an effect independently of the base rates that hold in a particular setting, whether it is important to consider the impact of base rates on the potential for prediction, or both. Effect sizes cannot be understood in a vacuum, and researchers have an obligation to consider the context or contexts in which an effect is to be understood.

References

Aaron, B., Kromrey, J. D., & Ferron, J. M. (1998, November). *Equating r-based and d-based effect-size indices: Problems with a commonly recommended formula.* Paper presented at the annual meeting of the Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED433353)

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed). Washington, DC: Author.

Anderson, C. A., Lindsay, J. L., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality? *Current Directions in Psychological Science, 8,* 3–9.

Baer, R. A., Wetter, M. W., & Berry, D. T. R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12,* 509–525.

Barratt, A., Wyer, P. C., Hatala, R., McGinn, T., Dans, A. L., Keitz, S., et al. (2004). Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction, and number needed to treat. *Canadian Medical Association Journal, 171,* 353–358.

Bartko, J. J. (1991). Measurement and reliability: Statistical thinking considerations. *Schizophrenia Bulletin, 17,* 483–489.

Beck, A. T., Brown, G., Berchick, R. J., Stewart, B. L., & Steer, R. A. (1990). Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients. *American Journal of Psychiatry, 147,* 190–195.

Beck, A. T., Steer, R. A., Kovacs, M., & Garrison, B. (1985). Hopelessness and eventual suicide: A 10-year prospective study of patients hospitalized with suicidal ideation. *American Journal of Psychiatry, 142,* 559–563.

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–598.

Blanchard, D. D., McGrath, R. E., Pogge, D. L., & Khadivi, A. (2003). A comparison of the PAI and MMPI-2 as predictors of faking bad. *Journal of Personality Assessment, 80,* 197–205.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41,* 687–699.

Carlson, M., Marcus-Newhall, A., & Miller, N. (1990). Effects of situational aggression cues: A quantitative review. *Journal of Personality and Social Psychology, 58,* 622–633.

Carroll, J. B. (1961). The nature of the data, or how to choose a correlation coefficient. *Psychometrika, 26,* 347–372.

Christensen, D., Hadzi-Pavlovic, D., & Jacomb, P. (1991). The psychometric differentiation of dementia from normal aging: A meta-analysis. *Psychological Assessment, 3,* 147–155.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Costner, H. L. (1965). Criteria for measures of association. *American Sociological Review, 30,* 341–353.

Davies, H. T. O., Crombie, I. K., & Tavakoli, M. (1998). Information in practice: When can odds ratios mislead? *British Medical Journal, 316,* 989–991.

Dawes, R. M. (1962). A note on base rates and psychometric efficiency. *Journal of Consulting Psychology, 26,* 422–424.

Dawes, R. M. (1993). Prediction of the future versus an understanding of the past: A basic asymmetry. *American Journal of Psychology, 106,* 1–24.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507–521.

Flegal, K. M., Carroll, M. D., Ogden, C. L., & Johnson, C. L. (2002). Prevalence and trends in obesity among U.S. adults, 1999–2000. *Journal of the American Medical Association, 288,* 1723–1727.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Forster, A. A., & Leckliter, I. N. (1994). The Halstead-Reitan Neuropsychological Test Battery for older children: The effects of age versus clinical status on test performance. *Developmental Neuropsychology, 10,* 299–312.

Funder, D. C., & Ozer, D. J. (1983). Behavior as a function of the situation. *Journal of Personality & Social Psychology, 44,* 107–112.

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Goodman, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data. *Journal of the American Statistical Association, 86,* 1085–1111.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3,* 339–353.

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

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58,* 78–79.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect-size displays. *Psychological Methods, 9,* 183–197.

Hunter, A. A. (1973). On the validity of measures of association: The nominal–nominal, two-by-two case. *American Journal of Sociology, 79,* 99–109.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). Thousand Oaks, CA: Sage.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1999). Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods, 4,* 257–271.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lester, D. (1995). The concentration of neurotransmitter metabolites in the cerebrospinal fluid of suicidal individuals: A meta-analysis. *Pharmacopsychiatry, 28,* 45–50.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oaks CA: Sage.

McKenna, M. C., Zevon, M. A., Corn, B., & Rounds, J. (1999). Psychosocial factors and the development of breast cancer: A meta-analysis. *Health Psychology, 18,* 520–531.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., et al. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56,* 128–165.

Narrow, W. E., Rae, D. S., Robins, L. N., & Regier, D. A. (2002). Revised prevalence estimates of mental disorders in the United States: Using a clinical significance criterion to reconcile 2 surveys' estimates. *Archives of General Psychiatry, 59,* 115–123.

Nunnally, J., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78,* 582–593.

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363.

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI-2. *Assessment, 1,* 227–237.

Rorer, L. G., Hoffman, P. J., Laforge, G. E., & Hsieh, K.-C. (1966). Optimum cutting scores to discriminate groups of unequal size and variance. *Journal of Applied Psychology, 50,* 153–164.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.

Rosnow, R. L., Rosenthal, R., & Rubin, D. R. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11,* 446–453.

Sackett, D. L., Deeks, J. J., & Altman, D. G. (1996). Down with odds ratios! *Evidence-Based Medicine, 1,* 164–166.

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Assessment, 8,* 448–467.

Schmidt, F. L. (1974). Probability and utility assumptions underlying use of the Strong Vocational Interest Blank. *Journal of Applied Psychology, 59,* 456–464.

Shrout, P. E., Spitzer, R. L., & Fleiss, J. L. (1987). Quantification of agreement in psychiatric diagnosis revisited. *Archives of General Psychiatry, 44,* 172–177.

Spitznagel, E. L., & Helzer, J. E. (1985). A proposed solution to the base rate problem in the kappa statistic. *Archives of General Psychiatry, 42,* 725–728.

Streiner, D. L. (2003). Diagnosing tests: Using and misusing diagnostic and screening tests. *Journal of Personality Assessment, 81,* 209–219.

Taylor, H. C., & Russell, J. T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23,* 565–578.

Thompson, B. (2002). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? *Journal of Counseling & Development, 80,* 64–71.

Thompson, I. M., Pauler, D. K., Goodman, P. J., Tangen, C. M.,

Lucia, M. S., Parnes, H. L., et al. (2004). Prevalence of prostate cancer among men with a prostate-specific antigen level less than or equal to 4.0 ng per milliliter. *New England Journal of Medicine, 350,* 2239–2246.

Torgersen, S., Kringlen, E., & Cramer, V. (2001). The prevalence of personality disorders in a community sample. *Archives of General Psychiatry, 58,* 590–596.

von Eye, A., & Mun, E. Y. (2003). Characteristics of measures for 2 × 2 tables. *Understanding Statistics, 2,* 243–266.

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

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374–378.

Appendix

Relationship Between Total and Pooled Standard Deviations

The following proof is applicable to both population and uncorrected sample variances but is presented in relation to the latter. The total variance of a set of scores can be partitioned into independent components as follows (e.g., see Cohen, 1988, p. 281, Equation 8.2.17):

$$S_Y^2 = S_{pooled}^2 + S_{\overline{Y}}^2.$$

In the case of two groups, this equation can be manipulated as follows:

$$\begin{aligned} S_Y^2 = S_{pooled}^2 + S_{\overline{Y}}^2 &= S_{pooled}^2 + E(\overline{Y}_{\cdot j}^2) - E(\overline{Y}_{\cdot j})^2 \\ &= S_{pooled}^2 + p_1\overline{Y}_{\cdot 1}^2 + p_2\overline{Y}_{\cdot 2}^2 - (p_1\overline{Y}_{\cdot 1} + p_2\overline{Y}_{\cdot 2})^2 \\ &= S_{pooled}^2 + p_1\overline{Y}_{\cdot 1}^2 + p_2\overline{Y}_{\cdot 2}^2 - p_1^2\overline{Y}_{\cdot 1}^2 - 2p_1p_2\overline{Y}_{\cdot 1}\overline{Y}_{\cdot 2} - p_2^2\overline{Y}_{\cdot 2}^2 \\ &= S_{pooled}^2 + (p_1 - p_1^2)\overline{Y}_{\cdot 1}^2 - 2p_1p_2\overline{Y}_{\cdot 1}\overline{Y}_{\cdot 2} + (p_2 - p_2^2)\overline{Y}_{\cdot 2}^2 \\ &= S_{pooled}^2 + p_1(1 - p_1)\overline{Y}_{\cdot 1}^2 - 2p_1p_2\overline{Y}_{\cdot 1}\overline{Y}_{\cdot 2} + p_2(1 - p_2)\overline{Y}_{\cdot 2}^2 \\ &= S_{pooled}^2 + p_1p_2(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})^2. \end{aligned}$$

The derivation of Equation 7 (the formula for the r_{pb} based on the pooled variance) from the more familiar Equation 5 (the formula for the r_{pb} based on the total variance of the quantitative variable) then proceeds as follows:

$$\begin{aligned} r_{pb} = \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{S_Y}\sqrt{p_1p_2} &= \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{\sqrt{S_{pooled}^2 + p_1p_2(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})^2}}\sqrt{p_1p_2} \\ &= \frac{(\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})}{\sqrt{\frac{S_{pooled}^2}{p_1p_2} + (\overline{Y}_{\cdot 1} - \overline{Y}_{\cdot 2})^2}}. \end{aligned}$$

Received January 31, 2005
Revision received May 3, 2006
Accepted August 24, 2006 ■