Psychological Methods
2008, Vol. 13, No. 1, 19–30

1082-989X/08/$12.00 DOI: 10.1037/1082-989X.13.1.19

A Probability-Based Measure of Effect Size: Robustness to Base Rates and Other Factors

John Ruscio
The College of New Jersey

Calculating and reporting appropriate measures of effect size are becoming standard practice in psychological research. One of the most common scenarios encountered involves the comparison of 2 groups, which includes research designs that are experimental (e.g., random assignment to treatment vs. placebo conditions) and nonexperimental (e.g., testing for gender differences). Familiar measures such as the standardized mean difference (d) or the point-biserial correlation (r_{pb}) characterize the magnitude of the difference between groups, but these effect size measures are sensitive to a number of additional influences. For example, R. E. McGrath and G. J. Meyer (2006) showed that r_{pb} is sensitive to sample base rates, and extending their analysis to situations of unequal variances reveals that d is, too. The probability-based measure A, the nonparametric generalization of what K. O. McGraw and S. P. Wong (1992) called the common language effect size statistic, is insensitive to base rates and more robust to several other factors (e.g., extreme scores, nonlinear transformations). In addition to its excellent generalizability across contexts, A is easy to understand and can be obtained from standard computer output or through simple hand calculations.

Keywords: effect size, nonparametric statistics, base rates, homogeneity of variance, independent groups

Calculating and reporting measures of effect size can assist researchers in many ways, such as by distinguishing statistical and practical significance (Kirk, 1996), cumulating and contrasting results via meta-analysis (Hunter & Schmidt, 2004), and estimating statistical power to plan studies (Cohen, 1988). Consequently, it is recommended that researchers routinely report appropriate measures of effect size (e.g., American Psychological Association, 2001; Wilkinson & the APA Task Force on Statistical Inference, 1999). One of the most common scenarios encountered in research is the comparison of two groups, which includes designs that are experimental (e.g., random assignment to treatment vs. placebo conditions) and nonexperimental (e.g., testing for gender differences). The most familiar measures of effect size for this scenario are the standardized mean difference (d) and the point-biserial correlation (r_{pb}). Although formulas exist to convert between d and r_{pb}, McGrath and Meyer (2006) demonstrated that these measures sometimes prompt different conclusions. In particular, McGrath and Meyer showed that whereas the value of d is unaffected by two groups' relative sample sizes, or *base rates*,[1] the value of r_{pb} attains a maximum value with equal-sized groups and declines as sample sizes diverge.

McGrath and Meyer's (2006) extensive analysis draws attention to an important issue, and they thoughtfully discussed the implications of this differential sensitivity to base rates for selecting an appropriate effect size measure. The insensitivity of d to base rates, however, requires one of the standard parametric assumptions: equal variances. As will be shown, d is sensitive to base rates when population variances are not equal. Researchers who would like to use

I thank Dan Phillips and Walter Kaczetow for the thoughtful comments and helpful suggestions they provided in response to drafts of this article.

Correspondence concerning this article should be addressed to John Ruscio, Department of Psychology, The College of New Jersey, P.O. Box 7718, Ewing, NJ 08628. E-mail: ruscio@tcnj.edu

[1] Unless otherwise specified, the term *base rates* is used to refer to the proportions of cases in two groups in a sample of data, not the relative sizes of the two populations from which these samples are obtained. For some populations, such as those corresponding to all those who could receive one treatment versus another for a certain condition, the population sizes may be equal. However, populations distinguished by features such as the presence versus absence of a diagnosable condition may differ in size considerably, and this difference itself may vary across contexts (e.g., the base rates of the condition may differ across institutionalized and noninstitutionalized populations). The focus of the present article is on the influence of sample base rates on effect size measures. These base rates may be consistent or inconsistent with those of the relevant populations, and this issue is discussed later.

a measure that remains insensitive to base rates even with unequal variances might consider a nonparametric version of the common language effect size statistic (described in Wolfe & Hogg, 1971, and studied by McGraw & Wong, 1992), a probability-based measure that is easy to understand. In the present article, I extend McGrath and Meyer's examination of effect size measures to situations with unequal variances and include a probability-based measure along with d and r_{pb}. In addition to demonstrating its insensitivity to base rates, I discuss many other strengths and limitations of the probability-based measure to facilitate the selection of the most appropriate effect size measure(s) to compare two groups in a particular situation.

Defining and Calculating Effect Size Measures

Standardized Mean Difference

The standardized mean difference in the population is defined as follows:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{1}$$

where μ_1 and μ_2 are the population means and σ is the population standard deviation, which is assumed to be equal across the two populations. Whereas δ often is referred to as Cohen's d, the convention adopted here is to use Greek symbols for parameters and Roman letters for statistics. McGrath and Meyer (2006) noted that there are many ways to estimate δ using sample data, and they provided a helpful organization of these variants (see their Table 1, p. 388). In keeping with their Equation 2 (p. 387), the sample standardized mean difference is calculated as

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p}, \tag{2}$$

where $\bar{Y}_1$ and $\bar{Y}_2$ are the sample means and s_p is the pooled standard deviation; s_p, in turn, is calculated as $s_p = \sqrt{\frac{SS_1 + SS_2}{N}}$, where SS_1 and SS_2 are the sums of squares for each group and N is the total sample size. (Following McGrath & Meyer, 2006, sample sizes rather than degrees of freedom are used in the denominators for sample variances.) Equivalently, this can be expressed as $s_p = \sqrt{p_1 s_1^2 + p_2 s_2^2}$, where p_1 and p_2 are the base rates in the sample ($p_1 + p_2 = 1$) and s_1^2 and s_2^2 are the sample variances. This yields the following formula for d:

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{p_1 s_1^2 + p_2 s_2^2}}. \tag{3}$$

Shortly, the consequences of pooling sample variances when population variances differ will be explored. For now, it is important to note that d requires the assumption of homogeneous population variances.

This effect size measure is interpreted as the difference between two groups' means on the dependent variable Y relative to the variability on Y within groups, calculated as a pooled estimate of the within-groups standard deviation. Cohen (1988) emphasized that the practical importance of an effect depends on the context of the research and offered rules of thumb to characterize its magnitude: d values of .20, .50, and .80 represent small, medium, and large effect sizes.

Point-Biserial Correlation

In their Equation 7 (p. 389), McGrath and Meyer (2006) expressed the sample point-biserial correlation using the same notation that was used for d. With the pooled variance written out as in Equation 3, this is

$$r_{pb} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{p_1 s_1^2 + p_2 s_2^2}{p_1 p_2} + (\bar{Y}_1 - \bar{Y}_2)^2}}. \tag{4}$$

This effect size measure is interpreted as the correlation between group membership and scores on the dependent variable Y, which is an estimate of the parameter ρ_{pb}. Assuming equal-sized groups, Cohen's (1988) rules of thumb for characterizing the magnitude of an effect as small, medium, and large are r_{pb}s = .10, .24, and .37, respectively. More generally, the r_{pb} corresponding to d for any base rates p_1 and p_2 can be calculated using a conversion formula shown in Table 1.

Probability-Based Measure

Parametric. McGraw and Wong (1992) described the common language effect size statistic (*CL*) as an attempt to communicate effect size information in a more intuitive way. *CL* estimates the parameter $\Delta = \Pr(Y_1 > Y_2)$, or the probability that a randomly chosen member of Group 1 scores higher than a randomly chosen member of Group 2. Cliff (1993) called this an ordinal answer to an ordinal question, which he argued is often more consistent with the research question that motivates investigation than more traditional comparisons of mean differences across groups. For example, when one is comparing a treatment group with a control group, *CL* estimates the probability that someone who receives the treatment would fare better than someone who does not. This communicates an important finding in concepts and language that are easy to understand, even without formal statistical training (see Hsu, 2004, for additional discussion of the intuitive appeal and statistical merits of a probability-based measure of effect size). Following Wolfe and Hogg (1971), McGraw and Wong (1992) calculated *CL* as follows:

$$CL = \Phi\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2 + s_2^2}}\right), \tag{5}$$

Table 1
Calculating and Converting Measures of Effect Size for Two Groups

Calculation	Conversions assuming equal-sized groups	Conversions without assuming equal-sized groups
$d = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{p_1 s_1^2 + p_2 s_2^2}}$ (1.451)	$d = \frac{2r_{pb}}{\sqrt{(1 - r_{pb}^2)}}$ (1.330)	$d = \frac{r_{pb}}{\sqrt{p_1 p_2 (1 - r_{pb}^2)}}$ (1.451)
	$d = \sqrt{2} \times \Phi^{-1}(CL)$ (1.265)	$d = \sqrt{\frac{s_1^2 + s_2^2}{p_1 s_1^2 + p_2 s_2^2}} \times \Phi^{-1}(CL)$ (1.451)
$r_{pb} = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\frac{p_1 s_1^2 + p_2 s_2^2}{p_1 p_2} + (\bar{Y}_1 - \bar{Y}_2)^2}}$ (.554)	$r_{pb} = \frac{d}{\sqrt{d^2 + 4}}$ (.587)	$r_{pb} = \frac{d}{\sqrt{d^2 + \frac{1}{p_1 p_2}}}$ (.554)
	$r_{pb} = \frac{\sqrt{2}\Phi^{-1}(CL)}{\sqrt{2[\Phi^{-1}(CL)]^2 + 4}}$ (.534)	$r_{pb} = \frac{\sqrt{\frac{s_1^2 + s_2^2}{p_1 s_1^2 + p_2 s_2^2}}\Phi^{-1}(CL)}{\sqrt{\left(\sqrt{\frac{s_1^2 + s_2^2}{p_1 s_1^2 + p_2 s_2^2}}\Phi^{-1}(CL)\right)^2 + \frac{1}{p_1 p_2}}}$ (.554)
$CL = \Phi\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_1^2 + s_2^2}}\right)$ (.814)	$CL = \Phi\left(\frac{d}{\sqrt{2}}\right)$ (.848)	$CL = \Phi\left(d\sqrt{\frac{p_1 s_1^2 + p_2 s_2^2}{s_1^2 + s_2^2}}\right)$ (.814)
	$CL = \Phi\left(\frac{r_{pb}}{\sqrt{.5(1 - r_{pb}^2)}}\right)$ (.826)	$CL = \Phi\left(r_{pb}\sqrt{\frac{\frac{p_1 s_1^2 + p_2 s_2^2}{p_1 p_2} + (\bar{Y}_1 - \bar{Y}_2)^2}{s_1^2 + s_2^2}}\right)$ (.814)

Note. d = standardized mean difference; r_{pb} = point-biserial correlation; CL = common language effect size statistic, whose nonparametric generalization is A. Converting between any of these measures requires the assumptions of normality and equal variances. $\bar{Y}_1$ and $\bar{Y}_2$ are the means, s_1^2 and s_2^2 the variances, and p_1 and p_2 the base rates for Groups 1 and 2. Φ is the normal cumulative distribution function; Φ^{-1} is the inverse normal cumulative distribution function. Values in parentheses beneath each equation were calculated for a data set with $\bar{Y}_1 = 2$, $\bar{Y}_2 = 0$, $s_1^2 = 4$, $s_2^2 = 1$, $p_1 = .30$, and $p_2 = .70$ to illustrate the influence of unequal variances and base rates on the conversions.

where Φ is the normal cumulative distribution function. As noted by McGraw and Wong, this technique requires the parametric assumptions of normality within groups and equal variances. When these assumptions are satisfied and groups are of equal size, the levels of d or r_{pb} representing small, medium, and large effect sizes are equivalent to $CL = .56$, .64, and .71, respectively (see Table 1 for conversion formulas).

Nonparametric. Cliff (1993) noted that ordinal statistics are more robust than parametric statistics to violations of the usual parametric assumptions when comparing mean differences, and many other researchers have advocated the use of a nonparametric technique to estimate $\Pr(Y_1 > Y_2)$ (Delaney & Vargha, 2002; Grissom, 1994; Grissom & Kim, 2001; Hsu, 2004; Vargha & Delaney, 2000). Allowing for the possibility of tied scores in a sample of data, Δ can be estimated using A, which is calculated as follows (Delaney & Vargha, 2002):

$$A = [\#(Y_1 > Y_2) + .5\#(Y_1 = Y_2)]/n_1 n_2, \quad (6)$$

where # is the count function.[2] In other words, one simply makes all pairwise comparisons between members of Group

[2] As noted earlier, in the present article, I use Greek symbols to denote parameters and Roman letters to denote statistics. With regard to the probability-based effect size measure, the parameter Δ is estimated using the parametric statistic CL or the nonparametric statistic A. Others have used different notation to refer to related measures. Delaney and Vargha (2002) defined the parameter $A_{12} = (\delta + 1)/2$, where $\delta = \Pr(Y_1 > Y_2) - \Pr(Y_2 > Y_1)$. Because $\Delta = \Pr(Y_1 > Y_2)$, Delaney and Vargha's $\delta = \Delta - (1 - \Delta) = 2\Delta - 1$ and $\Delta = (\delta + 1)/2 = A_{12}$. Hsu (2004) adopted the same notation as Delaney and Vargha; Grissom and Kim (2001) used the notation PS (for *probability of superiority*) to refer to an estimator of $\Pr(Y_1 > Y_2)$; and Cliff (1993) used δ and d to refer to what are symbolized here as Δ and A, respectively.

1 and members of Group 2, tallying the number of times that the former scores higher than the latter (or incrementing by 0.5 if they are tied), and divides by the total number of comparisons that were made.

A is related closely to several other statistics that require only ordinal data to estimate the difference between two groups, including the familiar Wilcoxon Rank Sum and Mann–Whitney *U* nonparametric test statistics as well as the area under a receiver operating characteristic (ROC) curve calculated using the trapezoidal method (Hanley & McNeil, 1982). Software can be used to obtain *A* with any of these procedures. SPSS reports the (trapezoidal) area under an ROC curve, which equals *A*, and Mann–Whitney $U = \#(Y_1 < Y_2) + .5\#(Y_1 = Y_2)$, in which case *A* can be calculated as

$$A = \frac{n_1 n_2 - U}{n_1 n_2}. \quad (7)$$

SPSS also reports the Wilcoxon test statistic as W_m, which can be converted to *U* (to calculate *A* as shown in Equation 7) as follows:

$$U = W_m - [n_s(n_s + 1)]/2, \quad (8)$$

where n_s is the smaller of the two sample sizes. Provided that one verifies how they are calculated, the nonparametric test statistics from other software can be used. For example, the R function for the Wilcoxon test reports $W = n_1 n_2 - U$, which equals the numerator in Equation 7.

Converting One Measure to Another

Rice and Harris (2005) presented tables of equivalent *d*, r_{pb}, and *CL* values under the assumptions of normality and equal variances, and they provided conversion formulas that apply in this case. Table 1 contains an expanded list of formulas for calculating these effect size measures or converting between them. The formulas in the first column show how to calculate each measure (these are Equations 3, 4, and 5). The formulas in the second column allow one to convert directly from one measure to another without knowledge of sample means, variances, or base rates, but they do so by relying on the assumption that group sizes are equal. These equations not only show the relationships between the measures in the simplest form (by assuming equal-sized groups), but they may be of use to a meta-analyst who needs to convert an effect size to a common metric but cannot determine the base rates in the original study. The conversion formulas in the third column do not require the assumption of equal base rates. These are provided primarily to show the relationship between measures in full detail; the meta-analyst who had sufficient information to use these conversion formulas (i.e., sample means, variances, and base rates) might find it easier to calculate the desired effect size measure rather than convert one that was reported.

All of the conversion formulas in Table 1 require the usual parametric assumptions of normality and equal variances. When these assumptions are satisfied, $A = CL$ and any of the formulas for converting to or from *CL* can be used to convert to or from *A*. When these assumptions are not satisfied, *A* and *CL* may differ in value; *A* would have to be calculated from the raw data.

Effect Sizes With Unequal Population Variances

Variance ratios in observed data ($s_1^2 : s_2^2$) vary tremendously. In a review of educational research, Keselman et al. (1998) found a mean variance ratio of 4:1, a median ratio of 2.25:1, and a maximum ratio of more than 560:1; Wilcox (2003) reported substantially larger values. Some of the observed difference in variances can be attributed to sampling error, but differences in population variances may be common nonetheless. To examine how much heterogeneity must be present in a sample of data to suggest that population variances differ, I submitted data sets with variance ratios ranging from 1.50 to 4.50 to O'Brien's (1981) test, which was recommended by Maxwell and Delaney (2004). At each variance ratio, the within-group sample size *n* began at 5 and was increased until the null hypothesis of population variance homogeneity was rejected at $\alpha = .05$. Data were generated as follows: Positive integers from 1 to $n + 1$ were transformed into percentiles, and these percentiles were converted into standard scores using the inverse normal cumulative distribution function; the first *n* values were retained. Scores for the second group were generated by multiplying scores in the first group by the square root of the desired variance ratio. Figure 1 plots the results of these

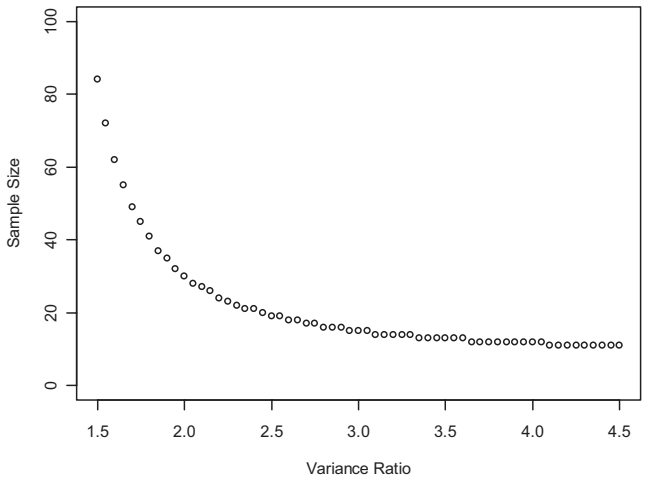

Figure 1. Minimum sample size (*n* per group) required to reject the null hypothesis of population homogeneity of variance at $\alpha = .05$ using O'Brien's (1981) test.

analyses. For the median variance ratio of 2.25:1 that Keselman et al. observed in educational research, $n = 23$ per group is required to reject the null hypothesis of population homogeneity of variance; at the mean ratio of 4:1, only $n = 12$ per group is required. It is not unusual for the ratio of standard deviations to exceed 2 in many areas of psychological research, in which case even small-sample studies often would support an inference of population heterogeneity of variance. It appears that the assumption of homogeneous variances that underlies the use of *d* as a measure of effect size often is not satisfied.

Under the condition of equal population variances for two groups, McGrath and Meyer (2006) found that r_{pb} was sensitive to base rates but that *d* was not. With unequal population variances, however, *d* becomes sensitive to base rates as well. This is because the pooling of sample variances involves weighting them by sample sizes. To the extent that a large proportion of a sample belongs to the group with the larger variance, this increases the pooled variance estimate and reduces *d*; the opposite occurs when a large proportion of a sample belongs to the group with the smaller variance. In contrast, the equivalence between *A* and the area under an ROC curve suggests that *A*, like the classification measures used to construct an ROC curve (sensitivity and specificity), should be independent of base rates. The definition and nonparametric calculation of *A* also suggest this independence: The probability that a randomly chosen member of one group scores higher than a randomly chosen member of the other group should not depend on how many people are in each group.

To examine these predictions, I calculated effect sizes across population variance ratios ($\sigma_1^2 : \sigma_2^2$) and population base rates ($\pi_1 + \pi_2 = 1$), with the absolute mean difference held constant at 2 by setting $\mu_1 = 2$ and $\mu_2 = 0$. Variance ratios of 1:1, 2:1, 4:1, 8:1, and 16:1 were used; σ_2 was held constant at 1, and σ_1 was set equal to the square root of the variance ratio. Base rates varied from $\pi_1 = .01$ to $\pi_1 = .99$ in increments of .001. Figure 2 plots population distributions for Group 1 (solid line) and Group 2 (dotted line) at each of these variance ratios for three illustrative base rates (π_1s = .10, .50, and .90). Any reasonable measure of effect size should yield smaller values when distributions overlap to a greater extent. In the present conditions, effect size should be inversely related to the variance ratio because the latter was a function of σ_1^2 alone; σ_2^2 was held constant, so increasing σ_1^2 to raise the variance ratio increased the extent to which the population distributions overlapped. To preserve the size of graphs within the figure, I used different scales for the *x*-axes as variance ratios increased. Whereas the absolute area of overlap between distributions decreases with increases in the variance ratio, the important point is that the proportion of overlap increases.

Figure 3 (left graph) shows the population standardized mean difference (δ) across conditions. Because the absolute mean difference was held constant at 2, δ achieved its maximum value of 2 when variances were smallest and equal ($\sigma_1^2 = \sigma_2^2 = 1$, represented by the solid line). With unequal variances, δ decreased in magnitude as a function of both the base rates and the variance ratio. Despite the fact that population means and variances remained constant within each variance ratio, δ decreased as the base rate for the group when the larger variance (π_1) increased. This decrease was more pronounced when the variance ratio was larger. Only when variances were equal was δ insensitive to base rates.

Figure 3 (middle graph) shows the population point-biserial correlation (ρ_{pb}) across conditions, with data points indicating the maximum value of ρ_{pb} attained for each variance ratio. The convex shape of the curve for equal variances (solid line) was explained well by McGrath and Meyer (2006). With equal variances, equal base rates yield the maximum value of ρ_{pb} because the variance of the dichotomous variable (the product of base rates) reaches a maximum when $\pi_1\pi_2 = .25$. As base rates diverge from one another, $\pi_1\pi_2$ decreases, and the constrained variance of the dichotomous variable reduces ρ_{pb}. As either group's base rate approaches 0, ρ_{pb} approaches 0 because group membership cannot correlate with scores on the dependent variable when all cases belong to the same group.

Because McGrath and Meyer (2006) only examined the influence of base rates when population variances were equal, they did not discover that ρ_{pb} varies asymmetrically across base rates when variances are unequal. In other words, for unequal variances, the maximum value of ρ_{pb} is attained with unequal base rates (see the solid data points plotted on the curves in Figure 3, middle graph). An important and perhaps counterintuitive implication is that statistical power is not necessarily maximized by studying equal-sized groups. When variances are unequal, one can achieve larger point-biserial correlations—and therefore greater statistical power—with groups of unequal but carefully chosen base rates. For a sample of data, neither the computed $t = \dfrac{r_{pb}\sqrt{N-2}}{\sqrt{1-r_{pb}^2}}$ nor the tabled value for $t_{\alpha,\, df = N-2}$ is dependent on base rates, so only the size of r_{pb} for a given *N* affects the statistical test's outcome.

Given estimates of s_1^2 and s_2^2 for a sample, one can estimate the base rates that would yield the maximum point-biserial correlation. Presuming that one cannot alter the mean difference between groups, the critical portion of Equation 4 appears in its denominator, $\dfrac{p_1 s_1^2 + p_2 s_2^2}{p_1 p_2}$. As this expression decreases, r_{pb} increases. The base rates yielding the minimum value for this expression are $p_1 = \dfrac{s_2}{s_1 + s_2}$ and $p_2 = \dfrac{s_1}{s_1 + s_2}$. For example, when $s_1^2 = 4$ and $s_2^2 = 1$, the

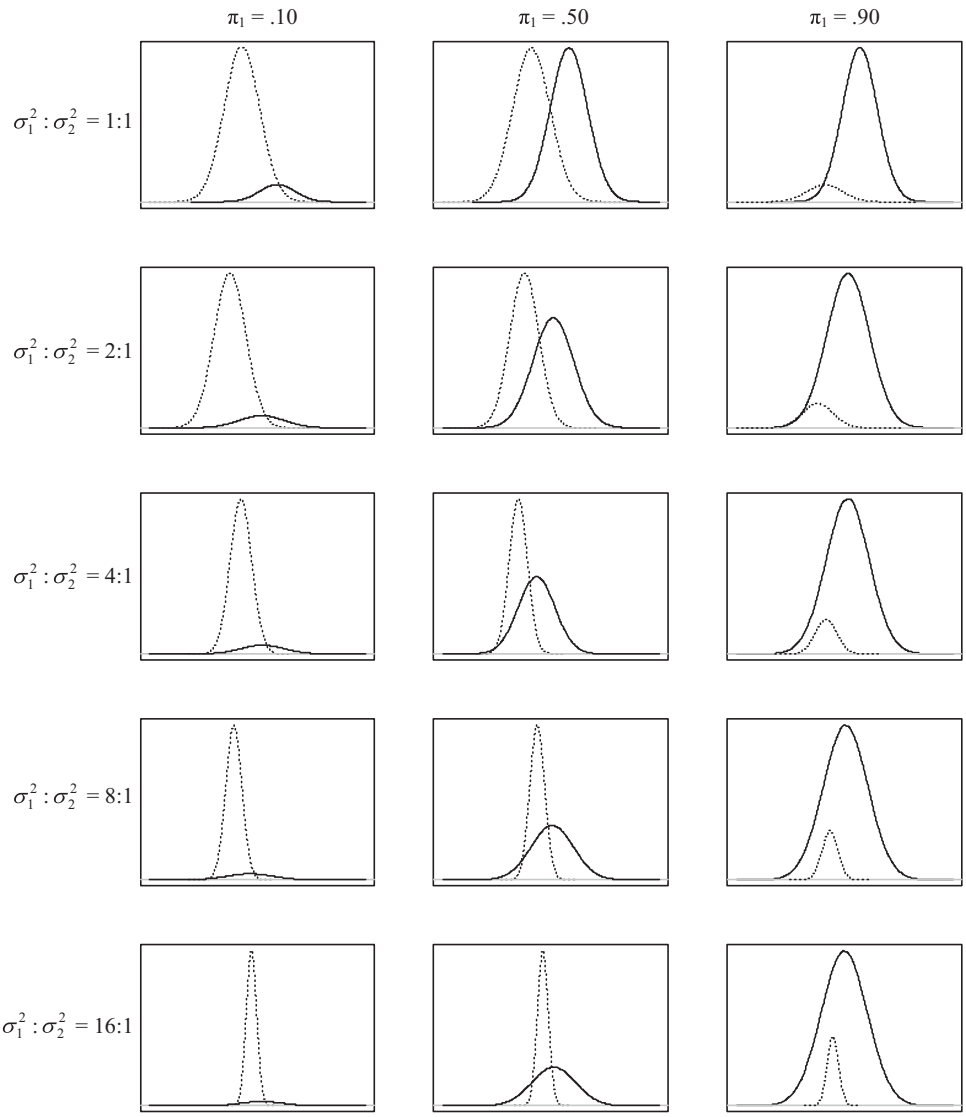

Figure 2. Population distributions as a function of base rates (labeled in columns) and variance ratios (labeled in rows). In each graph, the distribution for Group 1 is plotted as a solid line and the distribution for Group 2 is plotted as a dotted line. To preserve the size of graphs within the figure, different scales for the *x*-axes are used as variance ratios increase.

base rates yielding the maximum correlation would be $p_1 = 1/(2 + 1) = 0.33$ and $p_2 = 2/(2 + 1) = 0.67$.

Maximizing the size of r_{pb} (or d, or any other statistic) may not be the most appropriate goal in a given research context. For example, one might want to estimate the proportion of variance that an independent variable can explain in a dependent variable when group sizes are consistent with population base rates. Randomly sampling cases from the population of interest would yield a sample whose base rates would be representative of population values. Alternatively, if one uses a model that allows unequal population variances, the mean difference between two groups can be estimated with maximal precision by allocating more rather than fewer cases to the group with the larger variance. Freedman, Pisani, and Purves (1998) presented the standard error for the difference between means as $\sqrt{a^2 + b^2}$, where a and b are the standard errors of each mean. The base rates that minimize this standard error are the reverse of those indicated above—more cases are allocated to the group with the larger variance: $p_1 = \frac{s_1}{s_1 + s_2}$ and $p_2 = \frac{s_2}{s_1 + s_2}$. For the same case as above ($s_1^2 = 4$, $s_2^2 = 1$), the base rates that minimize $\sqrt{a^2 + b^2}$ would be $p_1 = 2/(1 + 2) = 0.67$ and $p_2 = 1/(1 + 2) = 0.33$. Clearly, one's research goals and

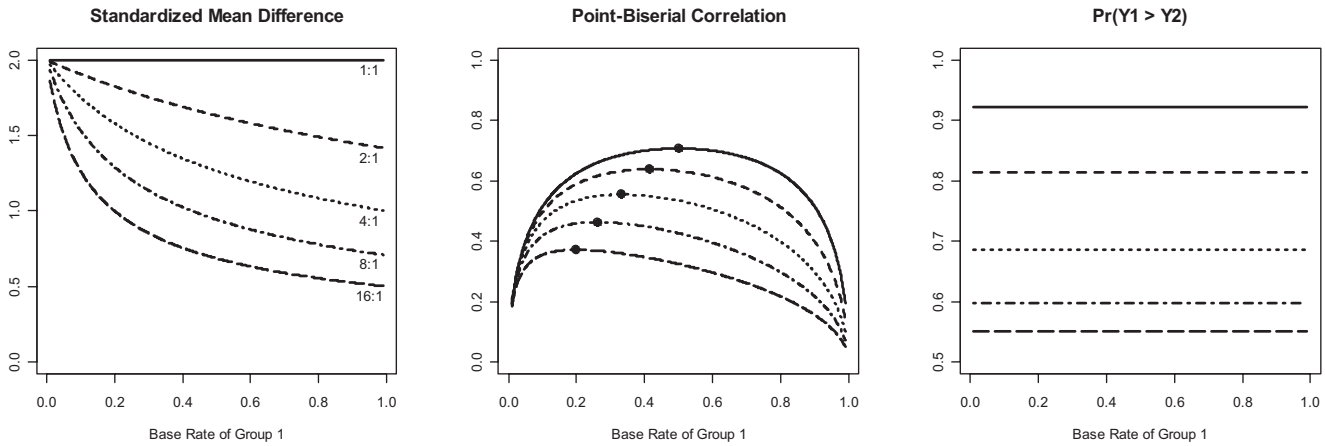

Figure 3. Effect sizes as a function of base rates (on the *x*-axis) and variance ratios (separate lines, labeled in the left graph). The maximum point-biserial correlation is plotted as a data point on each curve in the middle graph.

statistical model (including the assumptions that it requires) can influence the choice of an appropriate sampling strategy. For additional discussion of related points, see McClelland (1997) on how to optimize statistical power through the best allocation of cases to groups in a variety of research designs or Maxwell and Delaney (2004) on the challenges posed by comparing groups with heterogeneous population variances.

Figure 3 (right graph) shows the population probability-based measure of effect size (Δ) across base rates and variance ratios. Perhaps the most striking feature of these results is that whereas δ and ρ_{pb} differed as a function of both base rates and variance ratios, which confounds their influence at a constant mean difference, Δ neatly separates the influence of these two factors. As expected for each measure of effect size, an increase in the variance ratio yielded a smaller value of Δ. However, this influence on Δ did not depend on the base rates: The line for each variance ratio remained flat across π_1.[3]

An Empirical Illustration

So far, results have been obtained analytically using parameters. Do these idealized results generalize to conditions encountered in research? Each measure of effect size is subject to sampling error, and an empirical illustration helps to demonstrate the differential sensitivity to sample base rates even when calculated in samples of realistic size. Clark, Antony, Beck, Swinson, and Steer (2005) published data on the validation of the Clark–Beck Obsessive–Compulsive Inventory (CBOCI). A sample of 83 patients with a diagnosis of obsessive–compulsive disorder (OCD) scored higher ($M = 42.14$, $SD = 16.07$) than did a sample of 306 students ($M = 16.30$, $SD = 8.34$). Suppose new samples are drawn from normal distributions whose parameters are set equal to the means and standard deviations observed in Clark et al., varying the base rates from .02 to .98 in increments of .02. In other words, the total $N = 389$ and each population's μ and σ are held constant to examine what results these researchers might have found if they had sampled patients with OCD and students in different proportions. Following this plan, 1,000 samples were drawn at each base rate, and Figure 4 displays the results for each measure of effect size. Dark lines represent the mean results across all 1,000 samples at each base rate, lighter lines surround this by ± 1 standard error of the mean at each base rate, and horizontal dotted lines highlight the values of each measure at the base rates in Clark et al.'s data ($p_1 = 83/389 = .21$, $p_2 = 1 - .21 = .79$).

McGrath and Meyer (2006) concluded that "*d* can provide a better estimate [than r_{pb}] of the 'transportability' of an effect to an alternative context where the base rates differ" (p. 396). The illustrative analyses performed here, using a sample of data with a modest variance ratio (3.71:1), underscore the need to exercise caution with regard to the generalizability of *d* across base rates. Whereas patients with OCD and college students differed by $d = 2.47$ in Clark et al.'s (2005) study of the CBOCI, holding constant the mean and standard deviation of each group while varying their base rates yielded *d* values ranging from 1.63 to 3.02. The substantial difference in groups' variances suggests population heterogeneity, which calls into question the

[3] Whereas Δ separates the influence of unequal variances and base rates, the Mann–Whitney *U* or Wilcoxon tests of H_0 in which $\Delta = .50$ are not robust to heterogeneity of variance. Fligner and Policello (1981) presented a revised *U* test that is more robust, and a further improvement was identified by Delaney and Vargha (2002). Grissom and Kim (2005) also discussed this issue.

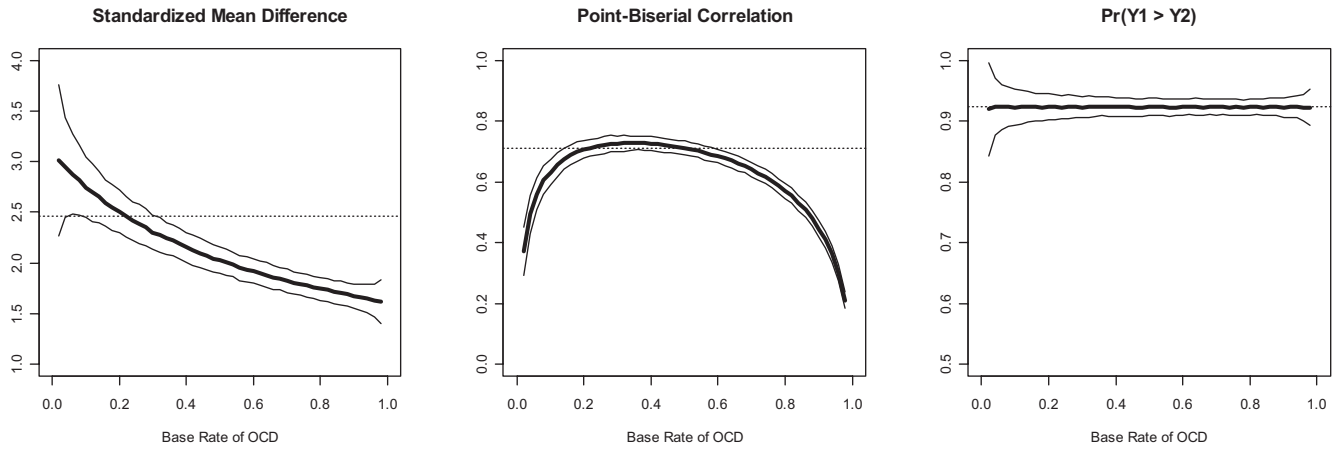

Figure 4. Effect sizes as a function of base rates (on the *x*-axis) in samples drawn from populations based on Clark et al.'s (2005) data. Dark lines show the mean values obtained for 1,000 samples at each base rate, and lighter lines represent the mean ±1 standard deviation at each base rate.

meaningfulness of *d*. This may not be an unusual situation, as psychological research often compares a group of "normal" participants with a group defined in such a way (e.g., by a diagnostic category) that individuals in the latter group would be expected to vary differently from one another than from individuals in the former group.

As expected, r_{pb} also varied widely (.21 to .73), yielding many values well below what was observed in the Clark et al. (2005) data (.71). In contrast, *A* yielded a highly consistent estimate of effect size (.92 to .93). To the extent that generalizability across base rates is an important factor when selecting a measure of effect size, *A* is superior not only to r_{pb} but to *d* as well.

In addition to demonstrating differential sensitivity to base rates, the results shown in Figure 4 also illustrate the sampling error of each effect size measure. For each measure, sampling error generally decreased as fewer cases were drawn from the population with a larger variance. For *d* and *A*, sampling error was especially pronounced as the base rates approached the extremes of 0 and 1. The variability shown in Figure 4 was observed rather than calculated, but formulas for the variance of each of these measures are available. Hunter and Schmidt (2004) described techniques for the meta-analysis of standardized mean differences or correlations, which included weighting each study's effect size measure by its estimated variance. Grissom and Kim (2001, p. 141) provided a formula to estimate the variance of *A*:

$$\hat{\sigma}_A^2 = [(1/n_1) + (1/n_2) + (1/n_1 n_2)]/12. \quad (9)$$

In addition to facilitating meta-analytic synthesis, the sampling error of a statistic is pertinent to the construction of confidence intervals (CIs). Along with the critical *t* value for a desired level of confidence (e.g., $\alpha = .05$, 2-tailed, is used for a 95% CI), the estimated variance could be used to construct a CI in the usual manner: $CI = A \pm (t_{\alpha,\, df = N - 2}) \times \sqrt{\hat{\sigma}_A^2}$. Alternatively, one might apply bootstrap methods to estimate σ_A^2 or to construct a confidence interval directly (see Efron & Tibshirani, 1993, for details). The optimal technique for estimating σ_A^2 warrants further study.

Artificial Dichotomization

Groups may be intrinsically discrete—whether naturally occurring (e.g., biological sexes) or experimentally manipulated (e.g., treatment vs. placebo)—or they may be created artificially through the dichotomization of a continuous independent variable (e.g., high vs. low scores on a personality scale to which a median split was applied). In the analyses presented above, it is presumed that an investigator has sampled from discrete populations, in which case the base rates reflect the relative sizes observed in a sample of data. When a continuous independent variable is dichotomized to produce groups for comparison, this usually reduces the magnitude of an effect by discarding information (Cohen, 1983; MacCallum, Zhang, Preacher, & Rucker, 2002). The extent to which the effect size observed in the resulting groups is attenuated depends on the threshold one applies (Hunter & Schmidt, 2004). Varying the threshold simultaneously alters the groups' relative sizes as well as the means and variances of the groups' scores on the dependent variable. As a result, not only is effect size usually weakened by artificial dichotomization, but this practice confounds the influence of several factors on measures of effect size.

To illustrate the complexity that this introduces, I drew samples of multivariate normal data ($N = 200{,}000$) from populations with $\rho = .10$, .30, .50, .70, and .90, and I

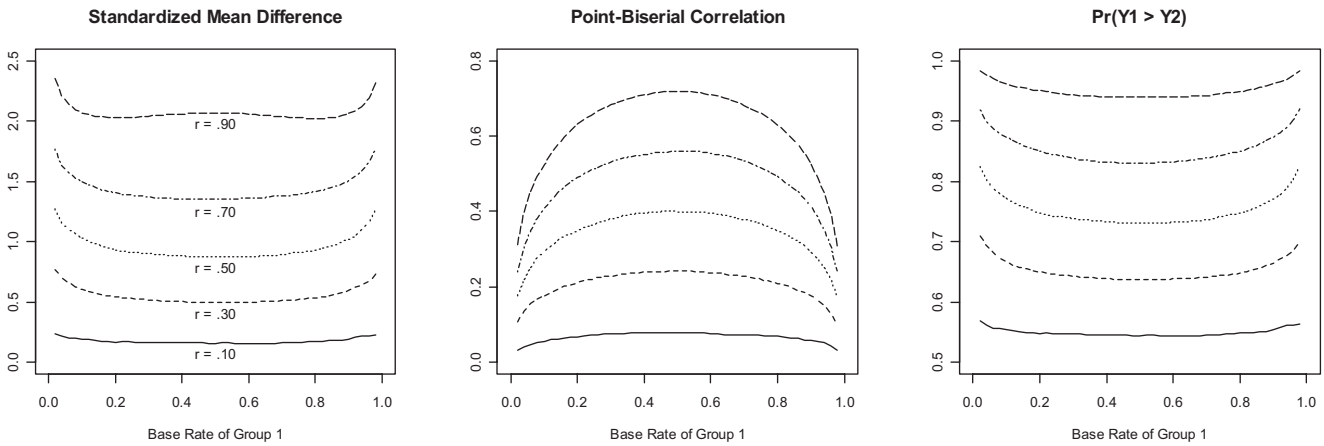

Figure 5. Effect sizes as a function of base rates (on the *x*-axis) and correlations prior to dichotomization (separate lines, labeled in the left graph). The original data were multivariate normal, with $N = 200{,}000$ at each correlation.

applied a series of thresholds to yield groups whose base rates ranged from .02 to .98 in increments of .02. For each sample and at each threshold, d, r_{pb}, and A were calculated; the results are shown in Figure 5. The curves for r_{pb} are similar to those appearing in Figure 3, where base rates varied due to differential sampling from discrete groups. The difference is that when thresholds were applied to dichotomize a continuous independent variable, correlations were reduced (as described by Cohen, 1983). The symmetric shape of these curves, which peaked when the threshold created equal-sized groups, is due to the multivariate normal distribution of the original independent and dependent variables. When data are not multivariate normal, the maximum value of r_{pb} will not necessarily be observed at equal base rates.

The value of d increased as base rates approached the limits of 0 or 1. As the threshold moved toward a very low (or high) value, it separated a small sample of low (or high) scores from a larger sample whose mean approached the value in the full sample. The more extreme the small sample's mean, the larger the value of d.[4] A similar pattern of results was observed for A. Once again, sensitivity to base rates under conditions of artificial dichotomization was due to the simultaneous changes in the central tendency and variability of the groups. As smaller groups of cases with relatively extreme scores are isolated from the remainder of the cases in a sample, this yields larger values of A because the extreme scores differ consistently from the others.

Neither d, r_{pb}, nor A is designed to estimate the population effect size ρ, which can be estimated without dichotomizing the independent variable and calculating r (Cohen, 1988). The methodological literature consistently cautions against artificially dichotomizing continuous variables, and the consequences of this practice on measures of effect size will not be considered further. All subsequent references to base rates once again presume that the groups being compared are intrinsically discrete.

Selecting the Most Appropriate Effect Size Measure(s)

Many factors are pertinent to determining how to conceptualize, measure, and report the size of a particular effect. Eight of these are discussed below and summarized in Table 2 to help researchers make informed choices between d, r_{pb}, and A. First, McGrath and Meyer (2006) thoughtfully considered the relative merits of d and r_{pb} in terms of the measures' differential sensitivity to base rates. However, in the present article, I have shown that depending on the extent of variance heterogeneity and the range across which base rates vary, either d or r_{pb} may be more sensitive to base rates. Thus, base rate sensitivity may not be very helpful in choosing between these measures. The probability-based, nonparametric measure A, in contrast, is insensitive to base rates. To the extent that insensitivity to the base rates in a particular sample of data is desirable, as when one wishes to generalize to other research or practical contexts, an investigator should give serious consideration to using A.

Second, McGrath and Meyer (2006) argued that d is more helpful than r_{pb} for understanding experimental or treatment effects. Nonetheless, the mean difference itself is not nec-

[4] Changes in variance within groups tend to diminish this effect on the value of d. As the small sample becomes more homogeneous, the large sample becomes more heterogeneous. Because the pooled variance term in the denominator of d weights each group's variance by its sample size, the net effect is an increase in the pooled variance that mitigates the effect of a larger mean difference in the numerator.

Table 2
Comparison of Effect Size Measures for Two Groups

Measure	*d*	r_{pb}	*A*
Definition	Standardized mean difference on dependent variable	Correlation between group membership and dependent variable	Probability that a randomly chosen member of Group 1 scores higher than a randomly chosen member of Group 2 on dependent variable
Sensitive to base rates	No, if equal variances; yes, if unequal variances	Yes	No
Relevance to understanding treatment effects	High if framed as a question of magnitude, low or moderate otherwise	Low	High if framed as an ordinal question, low or moderate otherwise
Connectivity to parametric models	Good	Good	Poor
Measurement scale requirements	Interval or ratio data	Interval or ratio data	Ordinal, interval, or ratio data
Ease of interpretation	Moderate	Difficult	Easy
Robustness to outliers	Poor	Poor	Good
Affected by order-preserving transformations	No, if linear; yes, if nonlinear	No, if linear; yes, if nonlinear	No
Robustness to violations of parametric assumptions	Poor	Poor	Good

Note. *d* = standardized mean difference; r_{pb} = point-biserial correlation; *A* = probability-based measure (nonparametric generalization of common language effect size statistic).

essarily important for certain practical purposes. As Cliff (1993) has argued, researchers and laypersons often pose an ordinal question (e.g., Do people who receive Treatment A tend to experience better outcomes than do people who receive Treatment B?) that can be answered most directly with an ordinal statistic. It is not difficult to construct plausible scenarios in which individuals tend to do better with Treatment A than with Treatment B, yet outliers or nonnormal distributions yield values of *d* that suggest the opposite. The choice between *d* and *A* depends on whether one prefers to quantify the magnitude of an effect in terms of the probability of one group's superiority over another or the standardized difference between the group means.

Third, McGrath and Meyer (2006) also noted some potential advantages of r_{pb} relative to *d*, including its more direct relationship to other statistical concepts (e.g., statistical power and the general linear model); the fact that it can be calculated across a wider range of study designs (there are correlation coefficients for dichotomous, ordinal, or continuous independent and dependent variables); and the fact that when base rates in a sample reflect those in the population, it provides a realistic sense for how well one variable predicts another. Because *A* is not a refinement of these conventional measures but an entirely different way of conceptualizing and expressing the difference between groups, it cannot offer the connectivity to parametric statistical models and inferences afforded by *d* or r_{pb}. Because parametric statistics predominate in primary studies and in meta-analyses, *A* might often serve most usefully as a supplement to—rather than a replacement for—either *d* or r_{pb}.

Fourth, even though *A* can be used when data are measured using interval or ratio scales, only ordinal-level measurement is required. When data are intrinsically rank ordered, neither *d* nor r_{pb} can be calculated. A related point is examined later: Reexpressing interval or ratio scale data using a nonlinear (but order-preserving) transformation can affect *d* or r_{pb} substantially, but it will not affect *A*.

Fifth, McGraw and Wong (1992) described the probability-based measure *CL* as one that would be simpler to understand or interpret than *d* or r_{pb}, especially when communicating results to individuals untrained in research design and statistical analysis. For example, consider how each measure expresses the same finding of Clark et al. (2005). Reporting *d* raises that question of which standard deviation to use. For example, one might state that the mean CBOCI score for OCD patients was $d = 2.47$ within-group standard deviation units above the mean for students, but when population variances are heterogeneous, the pooling of variances is difficult to justify. Instead, one might state that the mean score for OCD patients was either $d = 3.10$ student standard deviation units or $d = 1.61$ patient standard deviation units above the mean for students. Reporting *r*, one might state that group status (patient with OCD vs. student) correlated .71 with scores on the CBOCI. Reporting *A*, one

might state that the probability that a randomly chosen patient with OCD scored higher than a randomly chosen student was .92. For many people, the latter would be easier to understand. In advocating the use of *A* or equivalent measures, many researchers have emphasized its intuitive appeal (e.g., Cliff, 1993; Grissom & Kim, 2001; Hsu, 2004; Vargha & Delaney, 2000).

Sixth, *A* is much more robust to outliers than are *d* or r_{pb}. Both Cliff (1993) and Wilcox (2003) discussed the importance of robustness in great detail, and Wilcox demonstrated that many conventional statistics are highly sensitive to extreme scores. Because it is based on ordinal analysis, *A* is affected relatively little by outliers; *d* and r_{pb}, however, are highly susceptible to the influence of outliers. As an example inspired by Wilcox, consider samples of 1,000,000 drawn from normal versus mixed-normal populations with $\mu_1 = 1$, $\mu_2 = 0$. First, scores were drawn from normal populations, $\sigma = 1$. Next, scores were drawn from mixed-normal populations, $\sigma = 1$ with probability .90 and $\sigma = 10$ with probability .10. Although the mixed-normal distributions would be difficult to distinguish visually from their normal counterparts without careful inspection of the proportions of extreme scores in their tails, the distributions' variances differ substantially. For each normal distribution, $\sigma^2 = 1.00$, but for each mixed-normal distribution, $\sigma^2 = .90 \times 1^2 + .10 \times 10^2 = 10.90$. As a consequence, *d* was 70% smaller for the samples drawn from mixed-normal populations ($d = 0.30$) than for the samples drawn from normal populations ($d = 1.00$), and r_{pb} was 67% smaller (.15 vs. .45). By comparison, *A* was only 5% smaller (.72 vs. .76) and therefore considerably more robust to the influence of outliers.

Seventh, *A* is unaffected by order-preserving transformations. Whereas linear transformations (e.g., converting raw scores to standard scores) will not affect any measure considered here, nonlinear transformations will affect the values of *d* and r_{pb}. Nonlinear transformations often are used to help satisfy the assumptions of parametric statistical tests. For example, one might raise all scores to a common power to normalize distributions or equate their variances. Likewise, the act of measurement itself involves the use of a scale that may achieve no greater precision than a monotonic relation between observed and latent variables. In many cases, precision of measurement beyond an ordinal scale can be difficult to justify, and it may be illusory. Whether at the stage of measurement or data analysis, it is unfortunate that alternative measures achieving the same rank ordering of cases can yield very different values of *d* or r_{pb}. When this occurs, it is unclear which is the "correct" value or how well either would generalize to other contexts. As an example, consider samples of 1,000,000 drawn from chi-square distributions with 4 degrees of freedom vs. 2 degrees of freedom. Both distributions exhibited nontrivial levels of skew (γ_1s = 1.42 and 2.01, respectively) and kurtosis (γ_2s = 3.04 and 6.17, respectively), and their variances were unequal (variance ratio = 2.01). Effect sizes for these data were $d = .82$, $r_{pb} = .38$, and $A = .75$. How might effect size change after a nonlinear transformation was performed to satisfy parametric assumptions of normality ($\gamma_1 = 0$, $\gamma_2 = 0$) and equal variances? Trial and error revealed that raising all scores to the power of (1/2.244) reduced skew (γ_1s = .29 and .49, respectively) and kurtosis (γ_2s = −.05 and .01, respectively) and equated variances. Whereas *A* remained .75 after transforming the data, *d* increased by 16% to 0.95 and r_{pb} increased by 13% to .43. For *d* or r_{pb}, this raises the question of whether effect size should be estimated before or after performing transformations to meet the assumptions of parametric statistical tests.

Eighth, robustness to violations of parametric assumptions has been investigated for many measures of effect size. Hogarty and Kromrey (2001) performed a Monte Carlo study of the bias and variance of several effect size measures, including *d*, *CL*, and *A*. Of these three measures, *d* was most sensitive to violations of the parametric assumptions of normality and heterogeneity of variance, *CL* less so, and *A* yielded unbiased estimates of parameters and stable standard errors across normal and nonnormal population distributions as well as homogeneous and heterogeneous population variances. The use of trimmed means and Winsorized variances did not improve the robustness of *d* very much. More recently, Algina, Keselman, and Penfield (2005) reported that the use of trimmed means, Winsorized variances, and a bootstrap technique yielded CIs for a robust variant of *d* with good coverage probability for nonnormal population distributions. Whether a more robust version of *d* or r_{pb} can be achieved remains to be seen, but present evidence suggests that *A* is more satisfactory in terms of its operating characteristics.

Conclusions

A probability-based measure of effect size has much to recommend it, including conceptual and computational simplicity, communicative clarity, generalizability across research and real-world contexts that produce samples with different base rates, robustness to outliers, insensitivity to order-preserving data transformations, and robustness to violations of parametric assumptions. For those who prefer an ordinal statistic to help address ordinal questions, *A* should prove most useful; for those who prefer parametric statistical models, ordinal statistics do not provide the connectivity of *d* or r_{pb}. McGraw and Wong (1992) studied the *CL* measure, which possesses many desirable qualities, and demonstrated that it estimates Δ with some robustness to nonnormal population distributions, especially when population variances were equal. As anticipated by Cliff (1993), Hogarty and Kromrey (2001) found that a nonparametric generalization of *CL* was more robust to nonnormality and

variance heterogeneity than *CL* itself, retaining the elegance of a probability-based measure without requiring its parametric assumptions. Whether as a substitute for or supplement to more traditional measures such as *d* and r_{pb}, *A* merits a more prominent place among the psychological scientist's tools for expressing effect size.

References

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10,* 317–328.

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

Clark, D. A., Antony, M. M., Beck, A. T., Swinson, R. P., & Steer, R. A. (2005). Screening for obsessive and compulsive symptoms: Validation of the Clark–Beck Obsessive–Compulsive Inventory. *Psychological Assessment, 17,* 132–143.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114,* 494–509.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253.

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

Delaney, H. D., & Vargha, A. (2002). Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological Methods, 7,* 485–503.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* San Francisco: Chapman & Hall.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedures for the Behrens–Fisher problem. *Journal of the American Statistical Association, 76,* 162–168.

Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics* (3rd ed.). New York: Norton.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316.

Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods, 6,* 135–146.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* Mahwah, NJ: Erlbaum.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can we guess what they mean?* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

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

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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746–759.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d. Psychological Methods, 11,* 386–401.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365.

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin, 89,* 570–574.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d,* and *r. Law and Human Behavior, 29,* 615–620.

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the *CL* common language effect size statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25,* 101–132.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

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

Wolfe, D. A., & Hogg, R. V. (1971). On constructing statistics and reporting data. *The American Statistician, 25,* 27–30.

Received January 19, 2007
Revision received October 8, 2007
Accepted October 9, 2007 ■