r(equivalent): A Simple Effect Size Indicator

Psychological Methods
2003, Vol. 8, No. 4, 492–496

1082-989X/03/$12.00 DOI: 10.1037/1082-989X.8.4.492

$r_{equivalent}$: A Simple Effect Size Indicator

Robert Rosenthal
University of California, Riverside

Donald B. Rubin
Harvard University

The purpose of this article is to propose a simple effect size estimate (obtained from the sample size, *N,* and a *p* value) that can be used (a) in meta-analytic research where only sample sizes and *p* values have been reported by the original investigator, (b) where no generally accepted effect size estimate exists, or (c) where directly computed effect size estimates are likely to be misleading. This effect size estimate is called $r_{equivalent}$ because it equals the sample point-biserial correlation between the treatment indicator and an exactly normally distributed outcome in a two-treatment experiment with *N*/2 units in each group and the obtained *p* value. As part of placing $r_{equivalent}$ into a broader context, the authors also address limitations of $r_{equivalent}$.

Recent years have shown increasing dissatisfaction in psychology with the use of dichotomous decision-making based on significance tests and an increased recognition of the value of reporting effect sizes. Indeed, the report of the Task Force on Statistical Inference of the Board of Scientific Affairs of the American Psychological Association explicitly recommended that the primary results of any research should be presented as effect sizes, preferably with an accompanying confidence interval (CI; Wilkinson & the Task Force on Statistical Inference, 1999).

The purpose of the present article is to describe a simple procedure for obtaining an estimate of an effect size from a *p* value and the sample size. We call this effect size estimate $r_{equivalent}$. This procedure is especially appropriate when (a) in meta-analytic work, or in other reanalyses of others' studies, neither effect sizes nor significance test statistics (such as observed *t* values or *F* values) are provided, but only *p* values and sample sizes are reported; (b) no effect size estimate has been generally accepted for the data analytic procedures used; or (c) an effect size estimate can be computed directly from the data, but, because of small sample sizes or severe nonnormality, the estimates may be seriously misleading.

The basic idea is the following: Given the *p* value (i.e., obtained level of significance) from the actual study of a given size, what would be the point-biserial correlation ($r_{equivalent}$) if the *p* value had been obtained from the same-sized canonical (standard design) study? The more different the actual study is from the canonical study, the less relevant is the approximation using $r_{equivalent}$. If the actual study has the same form as the canonical study, then $r_{equivalent}$ is perfectly appropriate. Our choice for the canonical study is a two-group comparison of the means of a normally distributed outcome. There are other choices for a canonical study, but this choice seems more fundamental than any other in psychology.

Meta-Analytic Research in Which Only *p* Values Have Been Reported

In conducting meta-analyses we often find that only *p* values have been provided rather than effect size estimates or significance test statistics such as *t* or *Z,* or one degree of freedom *F* or χ^2 statistics. When those probability values are reported accurately (e.g., $p = .11$, $p = .02$, $p = .003$), we can use the method proposed here to obtain $r_{equivalent}$ from them and the sample sizes. When *p* values are reported only as $< .05$, $< .01$, and so on, we cannot get a unique value of $r_{equivalent}$, but we can set a lower bound, that is, the smallest possible value of $r_{equivalent}$, but not its upper bound, its largest possible value. The fact that in meta-analytic applications we can sometimes obtain

Robert Rosenthal, Department of Psychology, University of California, Riverside; Donald B. Rubin, Department of Statistics, Harvard University.

Order of authorship was determined alphabetically.

Correspondence concerning this article should be addressed to Robert Rosenthal, Department of Psychology, University of California, Riverside, California 92521-0426.

only lower bound values must be kept in mind, but such lower bound estimates of effect size are better than having no estimate at all, because in simple situations they may be regarded as conservative.

No Generally Accepted Effect Size Estimate Exists

Many effect size estimates have been described and have been widely used (e.g., Cohen, 1988; Fleiss, 1994; Rosenthal, 1991, 1994). However, in the recent history of statistical theory, considerably more work has been devoted to obtaining accurate p values than to developing indexes of effect size. Thus, there remain numerous statistical procedures for which no standard effect size estimate is recognized, for example, for many distribution-free or nonparametric procedures. What effect size estimate should we use, for example, when we have computed p values from Fisher's exact test, or from a sign test, a robust rank-order test, a Wilcoxon signed ranks test, a Mann–Whitney U test, or other permutation tests (Siegel & Castellan, 1988)? In such situations, $r_{\text{equivalent}}$ can be used, although as mentioned earlier, the more different the actual study and test statistic are from the canonical study, the less relevant is $r_{\text{equivalent}}$.

Directly Computed Effect Size Estimates Are Likely to Be Seriously Misleading

Consider a very small randomized experiment in which three animals are vaccinated and all survive, and three animals are not vaccinated and do not survive. The sample correlation between vaccination and survival for these six animals is 1.00. Because of the small sample size and the nonnormality of survival, the obtained sample correlation (r_{sample}) is probably a very misleading estimate of the population correlation. We can do better by computing an accurate probability for these six animals and then using the probability to compute a more appropriate effect size estimate, $r_{\text{equivalent}}$.

The Proposed Procedure

Our procedure yields $r_{\text{equivalent}}$ from an accurate one-tailed p value and sample size N by obtaining the value of t (with $df = N - 2$) associated with the one-tailed p value. When the p value we use to obtain the value of t is based on a contrast using more than two conditions, we obtain the value of t for $N - k$ *df,* where k is the number of conditions. The general principle is that we obtain the value of t with the degrees of freedom on which the p value is based. Although it is possible, in principle, to use two-tailed p values as well as one-tailed p values, we recommend consistent use of one-tailed p values to reduce ambiguity on this point. One-tailed p values in the "wrong" or unpredicted direction are recorded as $r_{\text{equivalent}}$ with a negative sign. We find these values of t quite readily from extended tables of *t,* from handheld calculators, or from computers. Once we have the t associated with the one-tailed p value and N, we compute $r_{\text{equivalent}}$ from

$$r_{\text{equivalent}} = \sqrt{\frac{t^2}{t^2 + (N-2)}}, \tag{1}$$

a well-known general relationship (Cohen, 1965; Rosenthal & Rosnow, 1991). When the p value we used to obtain the value of t was based on a contrast using more than two conditions, we replace the expression $(N - 2)$ in Equation 1 by the expression $(N - k)$, where k is the number of conditions. Even more generally, $N - 2$ is replaced by the degrees of freedom on which the p value is based.

The interpretation of $r_{\text{equivalent}}$ is that it is the sample point-biserial correlation we would have found in data yielding our obtained p value in a two-group, equal-group-size study with $N/2$ in each group, a study that we call the *canonical study.* That is, suppose we conducted a randomized experiment with $N/2$ assigned to the treatment condition and $N/2$ assigned to the control condition. Also suppose that the data are independently normally distributed in each condition with the same variance. Then, when the value of the t-test statistic is t with the obtained p value, the value of the point-biserial correlation between treatment condition and outcome is $r_{\text{equivalent}}$, given by Equation 1.

Although effect size estimates are available for omnibus, unfocused tests of significance (e.g., F with $df > 1$ in the numerator or χ^2 on $df > 1$), these are so much less specific and less interpretable that they should almost always be replaced by one or more single degree of freedom contrasts (Rosenthal, Rosnow, & Rubin, 2000).

CIs for $r_{\text{equivalent}}$

More research is needed to set appropriate CIs for $r_{\text{equivalent}}$. Until that research becomes available, however, we believe the usual procedure for forming CIs will work adequately for $r_{\text{equivalent}}$. Thus, a 95% CI

around the Fisher Z-transformed $r_{equivalent}$ can be found from the following equation:

$$95\%\ \text{CI} = Z_r \pm 1.96/\sqrt{N-3}. \quad (2)$$

The upper and lower limits computed from Equation 2 in units of Z_r are then transformed into their corresponding units of r using commonly available tables (e.g., Rosenthal & Rosnow, 1991, Table B.8).

Improving the Accuracy of r_{sample}: A Simple Example

Earlier we described a randomized experiment in which three vaccinated animals survived and three unvaccinated animals did not survive, yielding a sample correlation of 1.00 between being vaccinated and survival. We can obtain an accurate p value for these data from Fisher's exact test:

$$p = \frac{3!3!3!3!}{6!3!0!0!3!} = .05, \text{ one–tailed.}$$

Hence $p = .05$ and $N = 6$, so $t(4) = 2.13$, and from Equation 1 we find

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2 + (N-2)}} = \sqrt{\frac{(2.13)^2}{(2.13)^2 + (6-2)}} = .73,$$

a more realistic estimate of the population value of the correlation between vaccination and survival than the estimate of 1.00 based on the correlation in the sample, r_{sample}.

We now use Equation 2 to compute a 95% CI around the obtained $r_{equivalent}$. For $r_{equivalent} = .73$, we find $Z_r = .93$, so, with $N = 6$, the 95% CI around Z_r runs from $.93 - 1.96/\sqrt{3}$ to $.93 + 1.96/\sqrt{3}$ or from −0.20 to +2.06. Transforming our 95% CI for Z_r back to a 95% CI for $r_{equivalent}$ yields the interval from −.20 to .97. Had we tried to compute a 95% CI around the obtained value of r_{sample} (i.e., 1.00, with a Z_r value of $+\infty$), we would have found it to show no uncertainty at all, a result that is entirely unreasonable, because the population correlation is not known to be 1.00 based on those six data points.

$r_{equivalent}$ Versus r_{sample}

In what sense is $r_{equivalent}$ a more accurate estimate of the population correlation than is the sample correlation, r_{sample}? A formal answer to this question is based on the fact that r_{sample}, although approximately unbiased for the population correlation, in small samples is a poor (i.e., high-variance) estimate. For example, suppose that in the population 80% of vaccinated animals survive whereas only 20% of unvaccinated animals survive. That difference in survival rates is associated with a correlation between vaccination and survival of .60. If we repeated our experiment on three vaccinated and three unvaccinated animals over and over, we would often find r_{sample} of 1.00 even though we know the population correlation is only .60. If the population survival rate for vaccinated animals were 90%, whereas only 10% of unvaccinated animals survived, we would be even more likely to see r_{sample} values of 1.00, but our population value of r would still be far from 1.00; it would be .80. Even if 95% of vaccinated animals in the population survived, whereas only 5% of unvaccinated animals survived, we would still have a population correlation of only .90 while obtaining r_{sample} values of 1.00 most of the time.

Table 1 illustrates further that $r_{equivalent}$ based on exact p values behaves in an intuitively more realistic way than r_{sample} in small samples. Table 1 shows the results of seven hypothetical small-sample studies of the effects of treatment on primate survival with sample sizes ranging from 2 to 20. For each study, the p value reported is based on Fisher's exact test along with the associated $r_{equivalent}$ and the sample correlation, r_{sample}. As sample size, N, increases, the p value decreases, and $r_{equivalent}$ increases; however, r_{sample} never changes—it remains at 1.0.

Providing Effect Sizes Where None Are Currently Available: A Simple Example

Suppose that experts have ranked the performance of nine children on a reading ability measure where four randomly selected children were taught by a new method (treatment), and five children were taught by an old (control) method. All four of the treated children were ranked higher than any of the five control children, yielding an exact probability of .008, one-tailed Mann–Whitney U test (Siegel, 1956, p. 271). With $p = .008$ and $N = 9$, $t(7) = 3.16$, and from Equation 1 we find

$$r_{equivalent} = \sqrt{\frac{t^2}{t^2 + (N-2)}} = \sqrt{\frac{(3.16)^2}{(3.16)^2 + (9-2)}} = .77,$$

for which we find $Z_r = 1.02$ with the 95% CI extending from a Z_r of .22 to a Z_r of 1.82. Transforming our 95% CI into correlation units yields the interval from $r = .22$ to $r = .95$. Despite the limitations of

Table 1
Seven Studies Showing Sample Size, p Value, $r_{equivalent}$, and r_{sample}

Study	Condition	Results		N	One-tailed exact p	$r_{\text{equivalent}}$	r_{sample}
		Survive	Die				
1	Treatment	1	0	2	.50	.00	1.00
	Control	0	1				
2	Treatment	2	0	3	.33	.50	1.00
	Control	0	1				
3	Treatment	2	0	4	.17	.67	1.00
	Control	0	2				
4	Treatment	3	0	5	.10	.69	1.00
	Control	0	2				
5	Treatment	3	0	6	.050	.73	1.00
	Control	0	3				
6	Treatment	5	0	10	.0040	.78	1.00
	Control	0	5				
7	Treatment	10	0	20	.0000054	.82	1.00
	Control	0	10				

Note. $r_{\text{equivalent}}$ is computed from sample size and p value; r_{sample} is computed directly from observed data.

$r_{\text{equivalent}}$ that we describe shortly, it is useful to have a generally serviceable effect size estimate available where otherwise there would be none.

Choice of Effect Sizes

The same logic that leads to the use of $r_{\text{equivalent}}$ can be used to compute alternative indexes of effect size. For example, should we want to use Cohen's (1988) *d,* we would use Equation 3 to go from an obtained p value to its associated t and then find $d_{\text{equivalent}}$ from the following:

$$d_{\text{equivalent}} = \frac{2t}{\sqrt{N-2}}. \quad (3)$$

If our research consistently called for comparisons of only two groups, we could use $d_{\text{equivalent}}$ exclusively. In practice, however, effect sizes are often needed for contrasts based on more than two groups, for example, in computing linear trends or any other predicted pattern of three or more means with each pattern based on a single degree of freedom. In those situations it is less natural to use a two-group-based effect size indicator such as Cohen's *d,* Hedges's *g,* Glass's Δ, or related indexes. A limitation is that with more than two treatment conditions, $r_{\text{equivalent}}$ corresponds to r_{contrast}, not $r_{\text{effect size}}$—it is a fully partial correlation and therefore tends to overstate what might be viewed as the more natural effect size correlation; see Rosenthal et al. (2000) for discussion.

More detailed discussions of the use of various effect size indicators can be found in Rosenthal (1994) and in Rosenthal et al. (2000). The latter, in particular, describes a number of different correlational effect size indicators and gives reasons for often preferring them over various alternatives.

Nevertheless, the index $r_{\text{equivalent}}$ can be used in a wide variety of contexts beyond the simple contrasts computed among two or more treatment conditions. As long as a contrast is involved, comparisons among conditions leading to t tests or Z tests (or to F tests with one degree of freedom in the numerator, or chi-square tests on one degree of freedom) can all be used to compute $r_{\text{equivalent}}$. Examples include contrasts taken in the cells of any factorial design with only between factors, only repeated measures factors, or both. Also, we can use $r_{\text{equivalent}}$ in random effects or fixed effects analyses, for example, when the sampling units are conditions (random) or units nested within conditions (fixed) as in hierarchically nested designs. In meta-analytic applications, however, we typically do not have a choice of which to use to obtain $r_{\text{equivalent}}$. We are often able only to compute $r_{\text{equivalent}}$ for whatever type of analysis the original researcher reported; random effects in some cases, fixed in others.

The Meaning of $r_{\text{equivalent}}$ in Less Canonical Situations

As indicated in the previous discussion, because $r_{\text{equivalent}}$ is calculated from the obtained p value and

the associated degrees of freedom, it is not the same as effect sizes that use more information. For example, in the context of designs with one treatment with more than two levels, $r_{equivalent}$ is $r_{contrast}$ rather than $r_{effect\ size}$; in the context of factorial designs, again $r_{equivalent}$ is $r_{contrast}$ rather than the usually more appropriate $r_{effect\ size}$, or the even more generally appropriate $r_{effect\ size|NS}$, that is, given nonsubstantive factors—see Rosenthal et al. (2000, chap. 4); and in repeated measures designs, $r_{equivalent}$ treats all such designs as if they were intrinsically repeated measures designs (Rosenthal et al., 2000, chap. 5), which is clearly not always appropriate.

Similarly, in hierarchical designs with nested structure, or mixed and random effects models, there are choices of denominator error terms, and $r_{equivalent}$ corresponds only to the choice used to obtain the p value. And in analyses that produce a p value but are not adequately conceptualized as being a comparison of levels of a treatment (e.g., more than one numerator degree of freedom, more complex models), $r_{equivalent}$ may be deceptive, although we have not investigated such situations.

Limitations of $r_{equivalent}$

In closing, we want to emphasize that, although $r_{equivalent}$ is widely calculable, it is not a uniformly optimal procedure. It is not intended to be a kind of ubiquitous effect size indicator. It is, instead, designed specifically for those situations in which, first, the actual study is close in form to the canonical study, and, second, (a) the alternative is to have no effect size estimate at all (e.g., only sample sizes and p values are known for a study), (b) nonparametric procedures were used for which there are no currently accepted effect size indicators, or (c) sample sizes are so small or data so nonnormal that the directly computed effect sizes would be more misleading than the computed value of $r_{equivalent}$.

To conclude with a medical analogy: We think of $r_{equivalent}$ as a first-aid kit to be used for the time being until we can get to a highly sophisticated medical center. The medical center would be better, but it may be a long way away.

References

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

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

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research* (2nd ed.). New York: McGraw-Hill.

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

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

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

Received January 7, 2002
Revision received April 21, 2003
Accepted May 8, 2003 ■