Measures of effect size

William Roberts
Psychology Department
Thompson Rivers University
wlroberts@tru.ca

25 February 2013

In writing term papers, it is important to estimate the extent to which various possible causal factors are related to the outcome that interests you. This allows you to distinguish major (important) influences from minor (unimportant) ones. Moreover, if we can account for only a modest percentage of the total variability we observe, then there must be important limitations or problems in the methods used, or important causes have been omitted from the analysis. This paper briefly reviews various measures of effect size and describes how to calculate them from information usually given in a published report.

Tests of significance, in contrast to measures of effect size, only allow us to make an informed decision about whether an effect exists, that is, whether we will be able to replicate it – not how large or how important it is. When writing a paper, you will be able to definitively answer the question *Does this effect exist?* by looking for replication across studies. Therefore, don't worry about tests of significance: replication and effect size are much more important.

Unfortunately, articles published more than 5 or 10 years ago often omit measures of effect size. If you are reading such a paper, the simple techniques described below may be able to help you

1. The most important measures of effect size are *r*, the correlation coefficient, and

R^2, the proportion of total variability explained. The software available at http://www.tru.ca/faculty/wlroberts/#effect_size and at http://sourceforge.net/projects/behavioraldata/files/ can convert other measures of effect size (below) to R^2, so that R^2 can serve as a common metric across studies.

a. The correlation coefficient indexes the linear relation between two variables. If this is appropriate (some relations are curvilinear), then *r* indexes the strength of the relation between two variables in a world in which many factors influence each of them. (This is why it is difficult to infer causation from correlation.) Correlations range from -1 (low values on one variable perfectly predict high values on the other) through 0 (no correspondence) to +1 (each variable perfectly predicts the other, and both increase together from low to high). Thus the sign of a correlation (plus or minus) indicates the direction of the relation (positive or inverse) while the value indicates the strength of the relation.

b. It is often useful to consider a confidence interval for *r*. Because of sampling error, the sample correlation is only an estimate of the true, or population, correlation, and may differ from it substantially when the sample is not large. Given a correlation and a sample size, the calculator at http://www.tru.ca/faculty/wlroberts/r_confidence_interval.exe will return the 95% confidence interval, as well as the significance (*p*-value) of the correlation, and the critical values for *r* at various levels of alpha (.10, .05, .01, and .001) in a sample of the size given. One-tailed or two-tailed tests can be selected for these tests of significance. (For convenience, the calculator will insert a decimal point if the correlation is entered in digits only.)

c. R^2 (or Multiple R^2). In a study, the outcome measure varies from person to person. The proportion of this variability that we can predict or account for is R^2 (if there is one predictor) or multiple R^2 (if there are several predictors). Because it is a proportion, it ranges from 0 (no relation between the variables) to 1 (complete correspondence between the predictors and the outcome). You can convert a correlation coefficient to R^2 simply by squaring it. R^2 is usually given directly for multiple regression analyses and path analyses.

2. For analyses of variance (often abbreviated as ANOVA, MANOVA, ANCOVA), you should look for either ω^2 ("omega squared") or η^2 ("eta squared"). Both are equivalent to R^2. If neither is given, you can use the calculator at http://www.tru.ca/faculty/wlroberts/f_to_eta-square.exe. You will need to enter the two values for degrees of freedom and the value for F. For example, if an article reports $F(3,200) = 4.00$, $p < .01$, the first degree of freedom is 3, the second is 200, and the value of F is 4.00. The calculator returns $\eta^2 = .06$. Although the level of significance is high ($p < .01$) the effect is small (see below).

3. For either two-sample or matched-pairs[1] *t*-tests, you can calculate R^2 using http://www.tru.ca/faculty/wlroberts/t_to_rsquare.exe. You will need to enter the value of *t* and the degrees of freedom. For example, if an article reports $t(32) = 2.20$, $p < .05$, degrees of freedom = 32 and the value of *t* is 2.20. The square of the point-biserial correlation = proportion of variance in the outcome explained by group membership.
   a. The point-biserial correlation is appropriate if the two groups actually exist (e.g., girls and boys; treatment and control groups). If the groups have been formed by an arbitrary cut-point on an underlying continuous variable (e.g., pass-fail on an underlying continuum of achievement) then the biserial correlation should be used..

4. A measure of effect size frequently used in meta-analysis is *d*, the difference between the two means in units of standard deviations: $d = (\text{mean}_1 - \text{mean}_2) / \text{SD}$. You can convert *d* to *r* using http://www.tru.ca/faculty/wlroberts/d_to_rsquare.exe . For example, in their meta-analysis of the effects of divorce on children, Amato & Keith (1991) report effects ranging from $d = -.08$ to $d = -.26$, depending on the outcome assessed. That is, children from divorced families were, at worst, a quarter standard deviation below children from intact homes, a difference equivalent to a

[1] Gravetter & Wallnau (2008), *Statistics for the Behavioral Sciences,* p.349.

correlation of -.13. Divorce accounts for less than 2% of the variance.

5. For χ^2 (chi-squared) analyses, the measure of effect size is Cramer's *V* (sometimes referred to as "Cramer's ϕ"). If it is not given in the article, you can calculate it using http://www.tru.ca/faculty/wlroberts/chi_square_to_v.exe. You will need to enter the number of rows and columns (that is, the number of categories in each of the two variables analyzed), the total number of observations, and the value for χ^2.
   a. When we are dealing with a 2 x 2 table, $V = r$. For larger tables, we can interpret *V* as the average multiple correlation between columns and rows, which we might obtain by re-coding our data into (columns - 1) "dummy" (dichotomous) variables. In either case you may square it to obtain R^2.

6. The odds-ratio is a measure of effect size often used in medical research. It is indexing relative risk by comparing the probability of an outcome in each of two groups. For example, an odds-ratio of 2 (equivalent to an odds-ratio of ½) indicates that the proportion of an outcome in one group is twice that in the other group. Note that this is a *relative* effect size – it doesn't tell us how likely the outcome is – just that it is more likely in one group than in the other.
   a. You may convert an odds-ratio to R^2 by using http://www.tru.ca/faculty/wlroberts/odds_ratio_to_rsquare.exe. Use the values for ϕ (phi) when the groups actually exist (e.g., boys and girls; treatment and control; survivors and those who died). If the groups were formed arbitrarily from continuous variables, use the tetrachoric values.

How big is big?

The importance of a given effect size depends on the outcome being considered. In medical studies assessing survival, for example, even very small effect sizes are considered important because they represent differences in life or death. So context is critical when considering the importance of a given effect. However, for most psychological outcomes, it is common to follow the conventions proposed by Cohen (1988):

Correlations < .10 are so small they don't count; they are conceptually zero. Correlations from .10 to .30 are considered small. Those from .30 to .50 are moderate in size. Correlations greater than .50 are considered large.

Cutpoints for R^2 values are slightly higher, reflecting the fact that predictors in a multiple regression analysis are usually chosen for their size from a set of predictors. R^2 values less than .02 are conceptually zero. R^2 values from .02 to .13 are considered small; values from .13 to .26 are moderate in size; values greater than .26 are considered large.

Cohen suggests that for *d*, values less than .20 are so small that they don't count. Values from .20 to .50 indicate a small effect, values between .50 and .80 indicate a medium effect, and values greater than .80 are large. These cutpoints correspond to r^2 values of .01, .06, and .14.

Final thoughts: Effect size = 0?

To paraphrase Cohen & Cohen (1983, p. 57) there are times when the chilling thought crosses (or *should* cross) the mind of a reader that *none* of the correlations presented in a table departs from zero in the population. This is referred to as the omnibus null hypothesis, and while it is routinely assessed by such methods as the ANOVA (which assesses the rate of false positives across a set of comparisons between groups) and the MANOVA (which assesses the rate of false positives across a set of ANOVAs), published articles seldom assess the omnibus null hypothesis for sets of correlations.

If a published table of correlations seems to contain fewer significant findings than one might expect, you can assess the omnibus null hypothesis using the calculator at http://www.tru.ca/faculty/wlroberts/omnibus.exe. Enter the number of significant correlations; the total number of correlations tested (usually all the correlations in the table, but theoretically sensible subsets can be considered), and the alpha level used to judge significance (usually .05). The calculator will return the binomial probability of finding that number of significant correlations in a set that size, on the assumption that all the correlations are zero in the population. If the calculated probability is small, then you can reject the omnibus null hypothesis and conclude that some at least of the comparisons are not due to chance.

Remember that a certain number of false positives becomes a statistical certainty as the number of comparisons increases. Every study is likely to have at least a few. Replication is the only sure way – and the best – to identify patterns that actually exist in the population.

References

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N.J.: Erlbaum