CHAPTER 28

Evaluating Effect Size in Personality Research

Daniel J. Ozer

To provide an understanding of the origins and consequences of individual differences in personality, research commonly includes two or more variables and addresses the question: Are these variables related? Do persons high in agreeableness have more satisfying romantic relationships? Do those high in conscientiousness perform better at work? If such relations are thought to exist, one might ask the effect size question: What is the magnitude of the relationship between two variables (or groups of variables)? Questions of this kind are fundamental in personality research. Indeed, with the exception of univariate descriptive statistics and the statistical methods associated with hypothesis testing, most other quantitative methods used in personality research can be understood in the context of effect size estimation. Differences between means, correlation coefficients, regression coefficients, and the parameters of structural equation models are all example of effect size estimates. This chapter provides an account of effect size statistics as they are used in personality research. The primary focus is on the interpretation of these statistics, with estimation discussed only (and especially) insofar as interpretation depends on understanding how the effect size statistic has been generated.

What Do We Mean by "Effect Size"?

The term *effect size* has it origins in the statistical methods used to analyze experiments—"magnitude of experimental effect" (Friedman, 1968) is an equivalent label. Here, "effect" is used to refer to the impact of the causal, independent variable that is observed on the dependent variables. In the less than full rank analysis of variance model (Kirk, 1995, pp. 240), where Y_{ij} is the score of the ith person in the jth group, μ is the population mean (estimated by

the grand mean of Y), α_j is the *effect* of treatment j (estimated as the difference between the mean of the jth group and the grand mean), and ε_{ij} is a residual, error value (estimated by the difference between the individual i's score and the group's mean): $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$. Thus, one basic notion of effect size is $\alpha_j = \bar{y}_j - \bar{y}$—the difference between a group mean and a grand mean.

In the two group case, particularly where one is a treatment group and the other is a control, defining effect size as the difference between means, $\bar{y}_{Treatment} - \bar{y}_{Control}$, is an attractive alternative. In either case, the concept of effect size arises in the context of a bivariate relation between independent and dependent variables. Although adaptations to multivariate circumstances are possible, the straightforward simplicity of the bivariate case is considerably diminished as additional variables are included.

In personality research, it is frequently the case that quasi-experimental or correlational research designs are used, and in such designs the use of the term *effect* may be more analogical than literal. So, for example, if one examines the relationship between conscientiousness and job performance, one might refer to the magnitude of the obtained relation as the "effect size" despite the absence of a manipulated, independent variable. There is an implicit causal model that is untested in the research that provides a context for the assumption, and the estimate of effect size is contingent on the model. This broader, liberal use of causal language absent evidence in support of causality is adopted here, but it is important to remember that in this usage, as in personality research more generally, *effect size* may be a reference only to covariation between variables of interest.

Types of Effect Size Measures

This section describes two broad and widely applied types of effect size measures (raw and standardized) and briefly notes a third class of measures that are here labeled "criterion referenced" to parallel the meaning of that term in the context of testing.

Raw Measures of Effect Size

The most straightforward measure of effect size is simply the difference between two means. The interpretation of this measure depends on the units of measure and whether they are intrinsically meaningful. Most measures used in personality research are not directly and simply interpretable, and so raw measures of effect size have certain immediate disadvantages. Despite this drawback, raw effect size measures do have important uses because they are unaffected by sample standard deviations. When variables do have meaningful units, raw measures of effect size may be preferable.

When effect size is understood as degree of association, one occasionally useful effect size measure is the covariance. The sample covariance, computed as $s_{xy} = \sum(x - \bar{x})(y - \bar{y})/n$, is difficult to interpret because the meaning of the units in which it is expressed is entirely opaque. It is worth mentioning here only because the covariance is an important ingredient of the correlation coefficient that is the most widely used measure of effect in personality research. Unlike the correlation, the covariance has no upper or lower bound, and so the confidence interval surrounding sample estimates of population values are symmetric. For some purposes (e.g., in structural equation modeling), this property is sufficiently important to lead analysts to prefer the covariance rather than the correlation when inferential rather than descriptive goals are pursued.

Among the more common and useful raw measures of effect size is the unstandardized, or raw, regression coefficient: $b_{yx} = s_{xy}/s_x^2$, the ratio of the covariance to the variance of the predictor variable. This raw coefficient is the estimated change in y per unit change in x. In the special case where x is a dichotomy, with values of "0" and "1" a unit change in x is the difference between the two levels of the dichotomy, so the value b_{yx} is the mean difference in y between x units coded "1" and x units coded "0." For example, if a trait self-rating is obtained on a single 1–9 Likert scale and is regressed on participant gender (males arbitrarily assigned "1" and females arbitrarily assigned "0"), then the regression intercept is the mean of females on the self-rating, and the raw regression coefficient b_{yx} is the difference between females and males on the self-rating variable. So in this limited case, where x is a dichotomy and dummy coding is employed, the raw regression coefficient b_{yx} is simply the difference between two means.

When the predictor variable x is continuous, b_{yx} is the change in y associated with a unit change in x. It should be clear that the value of b_{yx}, which is expressed in the units of y, depends on the scaling of x. The change in weight associated with a unit change in height will depend on whether height is measured in inches, feet, or yards.

Standardized Measures of Effect Size

In the two group case, a raw effect size measure of some interest is the difference between group means: $\bar{y}_1 - \bar{y}_2$. As is generally the case, if the variable y is measured in interpretable units, then this raw effect size measure has much to recommend it. But when the units are not interpretable, this difference between means may be rescaled to reflect a standardized difference between means: $(\bar{y}_1 - \bar{y}_2)/s$. Depending on how the standard deviation, s, is calculated, this formula provides one of three different statistics. One approach developed for meta-analytic purposes comparing treatment and control groups (Glass, 1977; Glass, McGaw, & Smith, 1981) utilizes the estimate of the population standard deviation of the control group for s (i.e., sum of squared deviations from the control group mean divided by $n_{\text{control}} - 1$). Hedges g statistic (Hedges & Olkin, 1985) uses the estimate of the population standard deviation of both groups to obtain a pooled standard deviation estimate. The most frequently used effect size measure based on the standardized difference between means is Cohen's d, where s is the pooled sample standard deviations calculated with n rather than $n - 1$ in the formula for s (Cohen, 1977). To the extent that the variances of the two groups are the same, and as sample size gets large, these three measures of effect size converge; for most purposes, they may be given the same interpretation: The difference between sample means expressed in standard deviation units (i.e., the mean of either groups expressed as a z-score in the distribution of the other group).

Perhaps the most important and frequently used statistic in personality research is the Pearson coefficient of correlation, an effect size measure that expresses degree of association. Two formulas serve a definitional function for the correlation

$$r = \frac{\sum Z_x Z_y}{n} = \frac{s_{xy}}{s_x s_y}$$

The first expression shows that the correlation is the average cross-product of z-scores, or the covariance of a set of z-scores. In the second expression, the correlation is seen as the ratio of the covariance to the product of standard deviations. It is helpful to know that the maximum possible value of the covariance is the product of standard deviations, so the correlation is the percent of maximum possible covariance that is obtained in the data.

A frequent interpretation of r is based on its equivalence to β, the standardized regression coefficient, when there is but a single predictor variable in the regression equation. Thus, the correlation is the slope of the regression line relating scores on z_y to z_x.

Criterion-Referenced Measures of Effect Size

By a "criterion-referenced" measure of effect size, I mean a measure of effect that is tied to some meaningful standard outside the original units. It is a deliberate (but, I hope, useful) misnomer to refer to these as criterion-referenced measures of effect, because the measures of effect to be discussed here are in fact precisely those discussed above as raw measures of effect size. It is the units of measurement that are criterion referenced. This parallels the relation between standardized and raw effect-size measures—the scores are standardized and then effect size is determined.

A widely applicable kind of criterion-referenced effect size measure is based on "Percent of Maximum Possible" (Cohen, Cohen, Aiken & West, 1999), or POMP. Such scores are defined as 100 × (Observed score – Minimum score)/(Maximum score – Minimum score). This linear transformation of raw scores preserves several of the benefits of using raw scores while also providing several of the benefits associated with standardized scores. Consider a hypothetical case where the raw-regression coefficient based on POMP scores is .25. In this case, the difference or change on y associated with the difference or change on the entire possible range (maximum–minimum values) on x is the .25 value. That is, the maximum possible range of x is associated with one-quarter the maximum possible range on y. As in a standardized measure, this value is not dependent on the particular original units of the variables, but as in a raw measure, the results are not sample dependent, so that selection of

samples with broader or narrower ranges of observed scores should not alter the estimated effect size.

One concern that POMP scores suggest is the difficulty in knowing a small, medium, or large change or difference on a predictor variable x. A result like .25, above, is clear in relative terms: A difference of 1 unit on x is associated with a .25 unit difference on y; but it is quite possible that .25 on y is larger or more meaningful than a full unit on x. It is this phenomenon that makes Rosenthal's (1990) critique of standards used to assess effect size so powerful. There are instances where small or apparently minor treatments have dramatic effects (e.g., aspirin consumption on heart attacks). POMP measures are no more successful than other procedures in accommodating to certain kinds of interpretive weaknesses in our measures.

At least some kinds of interpretive difficulties can be resolved by importing a different kind of meaning into scores—meaning defined by the human observer. The just noticeable difference (jnd) of classical psychophysics provides a basis for interpreting the units of some scores, and methods for estimating scores in jnd units requires reference samples where the only information needed is the judge agreement, expressed as an average interjudge correlation and the validity of the judgment against the scale of interest (Ozer, 1993).

Other Methods of Characterizing Effect Size Measures

Broadly categorizing types of effect size measures in terms of the units used (raw, standardized, criterion referenced) is surely not the only approach, and there is benefit to considering alternatives. Vacha-Haase and Thompson (2004) suggest a different tripartite scheme—standardized mean differences, variance-accounted-for statistics, and corrected measures of effect size—to better estimate population values. All three of these types fall into my standardized or "norm-referenced" category, and it is certainly the case that if one wants to roughly divide the actual usage of effect size measures into roughly equal categories, the Vacha-Haase and Thompson approach does a better job than that employed above. Raw and criterion-referenced effect size measures are rarely used. Neither approach promotes consideration of partialled versus unpartialled estimates as a primary consideration, though this is certainly a matter of interpretative importance. Yet making distinctions between mean and correlation-based measures, or between sample values and estimates of population effects, or partialled and unpartialled effects, seems to place statistical matters first. Effect size estimation is at least in part a measurement problem, and if effect size values are to be understood, the first order of business is to understand the units used to express the effect. Following Cohen and colleagues (1999), I contend that reliance on standardized, norm-referenced units limits the potential accomplishments (e.g., any quantitative expression designed to describe or predict behavior or experience, using such units, will necessarily lack generality). Meaningful units of measure provide an opportunity for meaningful measures of effect.

The Interpretation of Effect Size Measures

Cohen (1962, 1977) originally proposed standards of "small" ($r = .1$, $d = .2$), "medium" ($r = .3$, $d = .5$), and "large" ($r = .5$, $d = .8$) to provide a rough interpretation of the magnitude of effects. In much of personality research, r^2 rather than r is used and interpreted as "percent of variance accounted for." Indeed, in its list of effect size measures, the APA (American Psychological Association, 2001) publication manual lists r^2 but not r (pp. 25–36) as an effect size measure. Rosenthal and Rubin (1982), Ozer (1985), Ahadi and Deiner (1989), and Abelson (1985), among others, have suggested that Cohen's standards may sell effects short and that squaring r exacerbates this tendency. It would seem that institutional guidelines and informal, traditional practice are each likely to lead to an underestimation of treatment effects and association among variables.

Abelson (1985) describes how intuitions about the meaning of "variance accounted for" can and frequently do create misconceptions about effect size. In an informal, but illustrative case, he shows how skill in batting accounts for substantially less than 1% of the variance in the outcome of a major league baseball at-bat. When Abelson polled colleagues to guess the percent of variance explained in an at-bat by batting skill, they overestimated the effect by a factor of 75! Either Abelson's colleagues are lacking knowledge about baseball (but then so

too are baseball fans, players, managers, and team owners who greatly value, each in their own currency, each increment in batting skill) or the variance-accounted-for metric seriously distorts judgments about effect size. Curiously, Abelson never fully resolves the question as to whether the variance-accounted-for measure or our lay intuition about this effect size is flawed. He does, however, suggest that when the natural causal structure of the world creates outcomes based on aggregated performance, the percent-variance-explained metric applied to a single unaggregated outcome will distort our understanding of effect size. Batting skill explains much more of the variance in a season's performance than in single at-bats. Likewise, extraversion may explain little variance in a measure of "new friends made today," but if there is any effect at all, the impact of extraversion on number of friends made in a year will be substantially larger.

Ahadi and Diener (1989) report two simulations showing that even when traits are the sole determinant of an outcome, if there are multiple causal traits, the correlation between any one trait and the outcome had an upper bound in the .45–.50 range. Interpreting the correlation between a trait and behavior of .21, the average effect size reported in the social and personality literature (Richard, Bond, & Stokes-Zoota, 2003) must be understood in this context. Given the myriad determinants of single behaviors under controlled circumstances, not to mention consequential outcomes reflecting years of effort and performance (e.g., measures of career success), a correlation of .25 suggests that the most potent, influential determinant of the outcome of interest has been successfully identified.

The correlation coefficient, r, and the coefficient of determination, r^2, are each a transparent nonlinear transformation of the other, and so at first glance it may seem that there could be little basis for preferring one or the other as a measure of effect size. Both r^2 and the absolute value of r range from 0.00 to 1.00, and so each statistic may appear to invite a percentage or proportion interpretation. Nearly all statistics textbooks and nearly all psychological researchers would endorse this practice insofar as it pertains to r^2 (as *per* the "coefficient of determination" name for this quantity and its interpretation as the percent of variance accounted for). But these same textbooks and researchers would deny that r has any meaningful or useful percentage interpretation. In this view, $r_{xy} = .2$ means that 4% of the variance in y is explained by x; and no meaningful interpretation of the relation between x and y invokes the quantity 20%. Some may go so far as to claim that half of a perfect correlation is not .5, but rather .707, because $r = .707$ yields a value of .50 for r^2. Why is the interpretation of r denied the privilege of a percentage interpretation that is granted to r^2?

The measure of variation with the property of additivity is the variance. This derives from the variance sum law, which states that the variance of a sum is the sum of all variances plus twice all covariances. Thus, r^2 is the percent of variance accounted for; $1 - r^2$ is percent of variance not accounted for (the variance of the residual values). One cannot form a parallel relation on the scale of correlation. That is, r can be thought of as the proportion of standard deviation accounted for, but the standard deviation of the residuals is $\sqrt{(1 - r^2)}$, not $1 - r$. This additivity on the scale of r^2 is especially important in multiple regression contexts. If predictor variables x and z are uncorrelated, then $R_{y.xz}^2 = r_{yx}^2 + r_{yz}^2$. This result enables the researcher to partition the total effect of a set of independent or predictor variables into parts attributable to each of the variables in the predictor set so that the sum of the effects of the parts are equal to the total effect. There is no such simple partitioning of effects on the scale of r, and it is this property of the scale of r^2 that accounts for the preference of textbooks and researchers for this metric.

Yet the validity of interpreting the partitioning of the squared multiple correlation, R^2, in terms of the r^2 (or squared, hierarchical semipartial correlations when predictors are correlated) values of the predictor variables is only rarely evaluated. Darlington (1990) develops an example that deeply undermines the use of the squared correlation metric for examining the relative importance of each variable. In this example, a nickel and a dime are (independently) tossed, and an outcome of a head results in winning the coin; if the outcome is a tail, nothing is won. In a single trial, there are four possible outcomes, each with a probability of .25: gains of nothing, 5¢, 10¢, and 15¢. Over multiple trials, it should be apparent that all the variance in winnings can be explained from the outcomes of the coin tosses, and the results for the dime coin are twice as important as the nickel coin results. Imagine the regres-

sion of winnings (possible values on any trial are 0, 5, 10, and 15) on the predictor variable "nickel result" (head coded 1, tail coded 0) and the predictor variable "dime result" (dummy coded in a parallel fashion). The expected results (nickel and dime outcomes independent and each of the four outcomes equally represented) are that the r^2 values for predicting winnings from the nickel and dime outcomes are .20 and .80, respectively. On the scale of r^2, it seems that dimes are four times more important than nickels in determining winnings, when it was clear, from the outset, that dimes should be only twice as important as nickels. Are regression statistics meaningless? Not when the appropriate statistic is used: The square roots of .2 and .8 are, respectively, .4472 and .8944. On the scale of r, dimes are exactly twice as important as nickels in determining winnings.

There is also a less widely known type of additivity associated with effect size on the scale of r. Consider the following thought experiment. Suppose that one wanted to generate a sample of observation pairs where the correlation between the pairs was exactly half of perfect. One might imagine creating one sample where the relation is perfectly present ($r = r^2 = 1.00$) and another sample, of the same size, where the relation is perfectly absent ($r = r^2 = 0.00$). What is the correlation in the combined, concatenated sample? The correlation, r, is .50, or half of perfect. This demonstration is depicted in Table 28.1. Under the specified conditions, additivity across variables accrues on the scale of r^2, but additivity over sampling units accrues on the scale of r. Although an explanation of additivity on the scale of r^2 is widely understood, additivity on the scale of r, as depicted in Table 28.1, has been a well-kept secret.

Final Comments

Measures of effect size figure prominently in the analysis of data collected in personality research. The reliability and validity of personality measures are most often evaluated with correlational measures of effect size, and personality structure is examined by factor analytic methods that rely on correlations among variables and provide correlations between variables and factors in the guise of factor loadings. But if any task is central to the enduring research program of personality psychology, it is the prediction of important life outcomes (Ozer & Benet-Martínez, 2006; Wiggins, 1973). Evaluating the success of this enterprise, in general, or the success of any particular prediction from a personality measure requires that measures of effect size be appropriately interpreted. Where it was once common practice to square a correlation coefficient and to interpret this diminished quantity uncritically on a simple percentage scale, it is now recognized that r as well as r^2 is needed to properly evaluate research progress (Rosenthal, 1990).

TABLE 28.1. The Additivity of Correlations over Sample Mixtures

	X	Y	
Sample A	1	1	
	1	3	
	2	2	r_{xy} = .00 in Sample A
	2	2	
	3	3	
	3	1	
			r_{xy} = .50 in combined sample
Sample B	1	1	
	1	1	
	2	2	r_{xy} = 1.00 in Sample B
	2	2	
	3	3	
	3	3	

Note. This demonstration of additivity of r over sample mixtures depends on the equality of means and standard deviations of each of the two variables across the samples.

Recommended Readings

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.

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

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.

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

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97*, 129–133.

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56*, 398–406.

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstances for POMP. *Multivariate Behavioral Research, 34*, 315–346.

Darlington, R. B. (1990). *Regression and linear models*. New York: McGraw-Hill.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70*, 245–251.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351–379.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

Kirk, R. (1995). *Experimental design: Procedures for the behavioral sciences*. Pacific Grove, CA: Brooks/Cole.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97*, 307–315.

Ozer, D. J. (1993). Classical psychophysics and the assessment of agreement and accuracy in judgments of personality. *Journal of Personality, 61*, 739–767.

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401–421.

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

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*, 775–777.

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

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51*, 473–481.

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment*. Reading MA: Addison-Wesley.