Educational Research Review 6 (2011) 135–147

Contents lists available at ScienceDirect

Educational Research Review

journal homepage: www.elsevier.com/locate/EDUREV

Review

Eta squared and partial eta squared as measures of effect size in educational research

John T.E. Richardson*

Institute of Educational Technology, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom

ARTICLE INFO

Article history:
Received 9 October 2010
Received in revised form
22 December 2010
Accepted 22 December 2010

Keywords:
Effect size
Eta squared
Partial eta squared

ABSTRACT

Eta squared measures the proportion of the total variance in a dependent variable that is associated with the membership of different groups defined by an independent variable. Partial eta squared is a similar measure in which the effects of other independent variables and interactions are partialled out. The development of these measures is described and their characteristics compared. In the past, the two measures have been confused in the research literature, partly because of a labelling error in the output produced by certain versions of the statistical package SPSS. Nowadays, partial eta squared is overwhelmingly cited as a measure of effect size in the educational research literature. Although there are good reasons for this, the interpretation of both measures needs to be undertaken with care. The paper concludes with a summary of the key characteristics of eta squared and partial eta squared.

Contents

* Tel.: +44 1908 858014; fax: +44 1908 654173.
E-mail address: J.T.E.Richardson@open.ac.uk

1747-938X/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.edurev.2010.12.001

1. Introduction

Most educational researchers appreciate, in abstract terms at least, that statements which describe the outcomes of tests of statistical inference need to be distinguished from statements which describe the importance of the relevant findings in theoretical or practical terms. Following a series of editorials and other articles by Thompson (1994, 1996, 1999, 2001) and the report of an American Psychological Association "task force" (Wilkinson & Task Force on Statistical Inference, 1999), many educational journals in North America require researchers to report measures of "effect size" as well as test statistics such as t or F (American Educational Research Association, 2006; Huberty, 2002). Journals elsewhere in the world tend to be less draconian, but authors are increasingly being encouraged to report such measures in their accounts of empirical research.

The term "effect" is most commonly associated with the use of different treatments in formal experimental research. However, nowadays it is commonly used by statisticians in the context of correlational research that examines the role of classification variables determined in advance of the research project rather than treatment variables that are under experimental control. For instance, the "effect" of gender on some measure of attainment is the difference between the scores obtained by male and female participants, or the "effect" of age might be the correlation between the participants' ages and their scores on the relevant attainment test (Sechrest & Yeaton, 1982). The measurement of effect size is thus an issue of relevance to all educational researchers, whether they are engaged in experimental or correlational research.

Fortunately there are good accounts of the development, interpretation and application of measures of effect size in education and cognate disciplines (e.g., Huberty, 2002; Olejnik & Algina, 2000; Richardson, 1996; Sechrest & Yeaton, 1982). Consequently, there is ample support for educational researchers wishing to develop expertise in the use of such measures. However, there is a lack of clarity in the literature concerning two particular measures called *eta squared* and *partial eta squared*. These were originally devised for measuring effect sizes in factorial designs, but nowadays they have a variety of other applications. Potentially, they are therefore of considerable importance in educational research. I recently published a paper in *Educational Research Review* that was intended as a tutorial overview of the development of measures of multivariate association (Richardson, 2007). In the present article, in the same spirit, I shall review the historical development of eta squared and partial eta squared and the issues involved in their application to assist other educational researchers and their students.

2. Eta squared

2.1. Defining eta squared

Most readers will be familiar with the linear correlation coefficient, Pearson r, which measures the degree of linear association between two variables (X and Y, say). Many will also be familiar with the fact that r^2, which is often termed the "coefficient of determination", is equal to the proportion of the total variation in one of the variables that can be predicted or explained on the basis of its linear relationship with the other variable (see, e.g., Hays, 1963, p. 505). Pearson (1903, 1905, pp. 10–11) devised analogous measures for a situation in which the different values of X represent various groups of participants. By way of example, Table 1 shows the results of an analysis of variance carried out on the data of a hypothetical study in which 100 12-year-old children, 50 boys and 50 girls, were given a test of reading ability.

The correlation ratio, η (eta), measures the degree of association between the two variables, the independent variable X (here, gender) and the dependent variable Y (reading performance). The square of the correlation ratio, η^2 (eta squared) is the differentiation ratio. It measures the proportion of the variation in Y that is associated with membership of the different groups defined by X. Using the notation in Table 1,

$$\eta^2 = \frac{\mathrm{SS(Between\,groups)}}{\mathrm{SS(Total)}} \tag{1}$$

where SS(Between groups) is the sum of squares for the effect of the independent variable and SS(Total) is the total sum of squares.

The correlation coefficient captures the linear relationship between X and Y, but the correlation ratio subsumes both the linear and nonlinear components of their association. If there are more than two groups and they are assigned numerical values in an arbitrary way, it makes no sense to talk about the "direction" of the association, and hence η is conventionally taken to be a positive quantity. Pearson (1905, p. 11) observed that $\eta \geq r$, with equality only when there is a linear relationship

Table 1
Results of hypothetical study comparing reading ability in 50 boys and 50 girls.

Source of variation	SS	df	MS	F	p
Between groups (gender)	300.00	1	300.00	30.00	.000
Within groups	980.00	98	10.00		
Total	1280.00	99			

Note: SS, sum of squares; MS, mean squares.

between the dependent variable and the numerical values assigned to the various groups defining the independent variable. Equivalently, the difference between the differentiation ratio and the coefficient of determination, $(\eta^2 - r^2)$, is an index of the deviation of the obtained regression curve from the least-squares regression line. The differentiation ratio is also equal to the squared multiple correlation coefficient obtained when the *X* variable is recoded as a series of independent dichotomous "dummy" variables (Cohen, 1969, p. 275). If there are merely two groups, η is equivalent to the point-biserial correlation coefficient between the independent variable and the dependent variable.

In the present example, $\eta^2 = \text{SS(Between groups)}/\text{SS(Total)} = 300.00/1280.00 = .234$ (to three decimal places). In other words, gender explains 23.4% of the variation among the children in terms of their reading ability. η can then be calculated as the square root of η^2 (in this case, .484). When using the statistical package SPSS to analyse genuine data obtained in a research design involving just one independent variable, the values of both η and η^2 can be readily obtained by using the MEANS procedure with the ANOVA option. In practice, however, η^2 is used far more often than η as a measure of effect size. Indeed, η^2 itself is sometimes incorrectly referred to as the correlation ratio (see, e.g., Hays, 1963, pp. 325, 547; Hedges & Olkin, 1985, pp. 101–102; Howell, 2002, p. 351; Kline, 2004, p. 99; 2009, p. 157; Winer, Brown, & Michels, 1991, pp. 209, 777, 780).

More generally, suppose that there are *k* groups constituting the *X* variable and that the total number of participants is *N*. From the definition given above and using the notation that is shown in Table 1,

$$\text{SS(Between groups)} = [\text{SS(Total)} \cdot \eta^2] \quad \text{and}$$
$$\text{SS(Within groups)} = [\text{SS(Total)} \cdot (1 - \eta^2)].$$

It follows that

$$\text{MS(Between groups)} = \frac{[\text{SS(Total)} \cdot \eta^2]}{(k-1)} \quad \text{and}$$
$$\text{MS(Within groups)} = \frac{[\text{SS(Total)} \cdot (1 - \eta^2)]}{(N-k)}.$$

Under the null hypothesis that there is no difference between the mean scores of the different groups, the two latter quantities are both independent estimates of the population variance in *Y*. It follows that the ratio between those estimates is distributed as the statistic *F* with $(k-1)$ and $(N-k)$ degrees of freedom (e.g., Diamond, 1959, p. 186; Hays, 1963, p. 548; McNemar, 1962, pp. 270–271). Algebraic manipulation yields a simpler version of this ratio:

$$F = \frac{[\eta^2 \cdot (N-k)]}{[(1-\eta^2)(k-1)]}. \tag{2}$$

For the example in Table 1, $F = [.234 \times (100-2)]/[(1-.234) \times (2-1)] = 22.932/0.768 = 30.00$ (using exact calculations rather than the rounded value of η^2 reported earlier).

2.2. Estimating eta squared

As originally defined, η and η^2 are statistics calculated from sample data. Like χ^2 (chi squared), they are exceptions to the modern convention that Roman letters denote sample statistics whereas Greek letters denote population parameters (see Halperin, Hartley, & Hoel, 1965). Kelley (1935) defined the population value of the differentiation ratio as the proportion of the total population variance in *Y* that was explained by membership of the various groups defined by *X*. If σ_Y^2 is the total population variance in *Y* and if σ^2 is the common variance within the groups (that is, the variance in *Y* that is *not* explained by membership of the various groups defined by *X*), then the population value of the differentiation ratio is equal to $(1 - \sigma^2/\sigma_Y^2)$. Following Pearson (1923), Kelley suggested that η derived from a sample tended to overestimate the population value of the correlation ratio.

An estimate is biased if it tends to be either systematically larger than the estimated value or systematically smaller than the estimated value. SS(Within groups)/$(N-k)$ is an unbiased estimate of σ^2, and SS(Total)/$(N-1)$ is an unbiased estimate of σ_Y^2. From this one might infer that an unbiased estimate of σ^2/σ_Y^2 might be obtained by taking the ratio between these two estimates, [SS(Within groups)/$(N-k)$]/[SS(Total)/$(N-1)$]. This can be rearranged to yield the formula [$(N-1)$·SS(Within groups)]/[$(N-k)$·SS(Total)]. Kelley therefore argued that an unbiased estimate of the population value of the differentiation ratio, which he called ε^2 (epsilon squared), was given by the complement of this formula:

$$\varepsilon^2 = 1 - \left\{ \frac{[(N-1) \cdot \text{SS(Within groups)}]}{[(N-k) \cdot \text{SS(Total)}]} \right\}. \tag{3}$$

It can be shown that $\eta^2 \geq \varepsilon^2$, with equality only when $\eta^2 = \varepsilon^2 = 1$. On the face of it, this seems to confirm that η tends to overestimate the population value of the correlation ratio. There was, however, a flaw in Kelley's reasoning. If *x* and *y* are two variables such that $x > 0$, then the expected value of the ratio y/x is greater than or equal to the ratio between their expected values, with equality only when *x* is actually a constant (see Kendall & Stuart, 1977, p. 242). Consequently, the right-hand term in Formula (3) tends to overestimate σ^2/σ_Y^2, and Formula (3) itself tends to *under*estimate the population value of the differentiation ratio.

Hays (1963, pp. 381–385) referred to the population differentiation ratio as ω^2 (omega squared), and he put forward a different estimate of this parameter which he called est. ω^2. However, this too was based on estimating the value of a fraction by means of inserting unbiased estimates of its numerator and denominator, and so it too tends to underestimate the population value of the differentiation ratio. Indeed, it can be shown that $\varepsilon^2 \geq$ est. ω^2, again with equality only when $\varepsilon^2 =$ est. $\omega^2 = 1$. Nowadays, it is generally recognised that both ε^2 and est. ω^2 are biased estimates of the population value of the differentiation ratio (e.g., Glass & Hakstian, 1969; Richardson, 1996; Winkler & Hays, 1975, p. 766), and η^2 is more widely used than either ε^2 or est. ω^2 as a measure of effect size in educational and social research.

2.3. Generalising to complex designs

The question arises how η^2 should be generalised to research designs in which there are two or more independent variables. For example, if there are two independent variables, *A* and *B*, then the total sum of squares can be divided into four components: SS(*A*), SS(*B*), SS($A \times B$), and SS(Within groups). Kennedy (1970) argued that the natural extension of Pearson's original conception as in Formula (1) would be to take the proportion of the total variation in the dependent variable that was associated with each individual effect or source of variation in the research design. For instance, the proportion of the total variation associated with the different groups defined by variable *A* is SS(*A*)/SS(Total). This is nowadays regarded as the classical interpretation of η^2 for factorial research designs (Olejnik & Algina, 2000).

A similar argument applies to research designs involving one or more covariates. An analysis of covariance identifies the linear relationship between the covariate(s) and the dependent variable *Y* and adjusts the values of the dependent variable to statistically control for that relationship. Both SS(Total) and SS(Within groups) are reduced as a result, and the difference between their reduced values is the revised SS(Between groups). Dividing the latter by the original SS(Total) expresses the proportion of the total variation associated with the different groups when the effects of the covariate(s) are statistically controlled (Olejnik & Algina, 2000). However, others have suggested that η^2 should be calculated by dividing the revised SS(Between groups) by the *reduced* SS(Total), effectively partialling out the effects of the covariate(s) (e.g., Tabachnick & Fidell, 1989, pp. 419–420; Winer et al., 1991, p. 780).

Finally, the classical interpretation of η^2 can be applied to within-subjects designs. If the independent variable is *A*, then SS(*A*) is one component of the within-subjects variation, and the residual within-subjects variation is used to test whether the effect of *A* is statistically significant. Even so, the size of the effect can still be expressed as a proportion of the total variation, and so once again $\eta^2 =$ SS(*A*)/SS(Total) (Kline, 2004, p. 116; Olejnik & Algina, 2000). However, one would not expect a within-subjects variable to explain any of the between-subjects variation. It could therefore be argued that η^2 should be calculated as a proportion of the within-subjects variation, effectively partialling out the between-subjects variation.

3. Partial eta squared

3.1. Defining partial eta squared

In Section 2.1, I showed that, in a one-way analysis of variance, the statistic *F* can be expressed as a function of η^2 (see Formula (2)). Cohen (1965) pointed out, conversely, that the corresponding values of η and η^2 could be calculated from the value of *F*. In the present notation, this yields the following formula:

$$\eta^2 = \frac{[F(k-1)]}{[F(k-1)+(N-k)]}. \tag{4}$$

For the example in Table 1, $\eta^2 = [30.00 \times (2-1)]/\{[30.00 \times (2-1)] + (100-2)\} = 30/(30+98) = 30/128 = .234$. In the special case when there are just two groups, η can be calculated from reported values of Student's *t* by means of the following formula (Richardson, 1996):

$$\eta^2 = \frac{t^2}{(t^2+N-2)}.$$

It follows that interested but sceptical readers can calculate values of η and η^2 for themselves from the inferential statistics that are reported in published research articles. Of course,

$$\begin{aligned} F &= \frac{\text{MS(Between groups)}}{\text{MS(Within groups)}} \\ &= \frac{[\text{SS(Between groups)}/(k-1)]}{[\text{SS(Within groups)}/(N-k)]}. \end{aligned}$$

Substituting the latter expression into Formula (4) yields the following:

$$\eta^2 = \frac{\text{SS(Between groups)}}{[\text{SS(Between groups)} + \text{SS(Within groups)}]}. \tag{5}$$

For research designs in which there is only one independent variable (as in Table 1), SS(Total) = SS(Between groups) + SS(Within groups). It follows that Formulae (1) and (5) are equivalent. However, this is not the case for research

designs with more than one independent variable. For example, suppose that there are two independent variables, *A* and *B*. When determining the value of η^2 associated with variable *A*, Formula (4) employs the *F* ratio for the effect of that variable, which is MS(*A*)/MS(Within groups). Formula (5) then takes the following form (see also Cohen, 1973; Olejnik & Algina, 2000):

$$\eta^2 = \frac{\mathrm{SS}(A)}{[\mathrm{SS}(A) + \mathrm{SS}(\text{Within groups})]}. \tag{6}$$

In other words, SS(*B*) and SS($A \times B$) are removed from SS(Total) before calculating the proportion of variation that is associated with the different groups defined by variable *A*.

As Cohen (1965) explained:

> It must again be stressed that *any* source of variation which yields an *F* ratio can equally yield η as an index of "how much" relationship there is between this source and the dependent variable, other nonerror sources of variation being partialed out, or η^2 as the proportion of the relevant sum of squares for which this source accounts. (p. 105, italics and spelling as in original)

Note that, like the classical versions of η and η^2, Cohen was proposing these two measures as statistics to be calculated from sample data. In this account, however, Cohen did not explain the rationale for preferring these measures of effect size over the classical versions or why, in particular, it was necessary or desirable to partial out the other nonerror sources of variation.

3.2. The standard deviation of the standardised population means

This explanation can be found elsewhere in Cohen's writings where he focused on the statistical power of published research: that is, the probability of correctly rejecting a false null hypothesis and thus detecting a genuine effect. To begin with, Cohen (1962) analysed each of the studies reported in a then recent volume of the *Journal of Abnormal and Social Psychology*. From the information contained in the relevant article, he calculated the probability that the researcher would have detected "small", "medium", and "large" effects, where these were operationally defined for different kinds of statistical test. He found that these probabilities were quite low (for instance, the odds of detecting a medium effect were less than 50:50), and he suggested that researchers needed to increase the statistical power of their studies (most obviously by using larger samples of participants).

In this case, effect size was discussed in terms of the properties of the populations from which the participants had been drawn, and consequently it was defined in terms of particular population parameters. For studies that had involved comparisons among several groups of participants, Cohen defined the population parameter *f* as the ratio between the standard deviation of the population means and the standard deviation within the populations. If the population mean for the *i*th group is μ_i and the grand mean of the population means is μ, the standard deviation of the population means is

$$\sigma_m = \sqrt{\left\{\frac{[\sum(\mu_i - \mu)^2]}{k}\right\}}.$$

Thus, $f = \sigma_m/\sigma$. Equivalently, *f* is the standard deviation of the population means when they have been standardised against the standard deviation within the populations.

Cohen suggested that, in this research design, small, medium, and large effects would be reflected in values of *f* equal to 0.10, 0.25, and 0.50, respectively. He commented: "These values are necessarily somewhat arbitrary, but were chosen so as to seem reasonable. The reader can render his own judgment as to their reasonableness. . ." (p. 146). In other words, Cohen specified these values *ex cathedra* rather than on the basis of any empirical evidence. Within a specific field, experienced researchers might have their own ideas about what would constitute small, medium and large effects that were different from Cohen's benchmarks.

Subsequently, Cohen (1969) devoted an entire book to the subject of statistical power analysis. Here, he pointed out that there was a relationship between *f* and the population correlation ratio, which he continued to designate by η (pp. 273–274). In particular, the total population variance in *Y*, σ_Y^2, is equal to $(\sigma^2 + \sigma_m^2)$. Consequently,

$$\eta^2 = \left[\frac{\sigma_m^2}{(\sigma^2 + \sigma_m^2)}\right].$$

It follows that

$$\eta^2 = \left[\frac{f^2}{(1+f^2)}\right] \quad \text{and}$$

$$\eta = \sqrt{\left[\frac{f^2}{(1+f^2)}\right]}.$$

Algebraic manipulation shows that, conversely,

$$f^2 = \left[\frac{\eta^2}{(1-\eta^2)}\right] \quad \text{and}$$

$$f = \sqrt{\left[\frac{\eta^2}{(1-\eta^2)}\right]}.$$

On this occasion, Cohen suggested that small, medium, and large effects would be reflected in values of f equal to 0.10, 0.25, and 0.40, respectively. These correspond to values of η of .100, .243, and .371, respectively, and to values of η^2 of .0099, .0588, and .1379, respectively (pp. 278–280).

3.3. Generalising to complex designs

Cohen (1969) then considered the application of f to factorial designs (pp. 358–360). If there are two independent variables, A and B, the population means defined by variable A are still standardised against the standard deviation within the populations, σ. Consequently,

> the definition of f. . .remains the same—the standard deviation of the. . .standardized means, where the standardization is by the common within (cell) population standard deviation. . . Thus, there is no need to adjust one's conception of f for a set of k means when one moves from the one-way analysis of variance. . .to the case where additional bases of partitioning of the data exist. . . It is, however, necessary to consider the interpretation of η^2.
>
> In factorial design[s], the total variance is made up not only of the within (cell) population variance and the variance of the means of the levels of the factor under study, but also the variances of the means of the other factor(s) and also of the interactions. Therefore, the variance base of η^2. . ., namely $\sigma^2 + \sigma_m^2$, is no longer the total variance, and the formulas involving η and η^2. . .require the reinterpretation of η as a *partial* correlation ratio, and η^2 as a proportion, not of the total variance, but of the total from which there has been excluded (partialed out) the variance due to the other factor(s) and interactions. (p. 359, italics in original)

This "reinterpretation" of η^2 meant that the relationship between f^2 and η^2, $\eta^2 = [f^2/(1+f^2)]$, remained true. As a result, the values of η^2 for the individual effects and interactions in the research design were not affected by the size of the other effects and interactions. Cohen maintained that his proposed operational definitions of small, medium, and large effects in terms of f (and by implication in terms of η^2, too) had their usual meaning (p. 360).

Cohen applied this approach to designs with one or more covariates (pp. 372–373). As mentioned above, an analysis of covariance identifies the linear relationship between the covariate(s) and the dependent variable Y and adjusts the values of the dependent variable to statistically control for that relationship. The effect size index f is still equal to σ_m/σ, but σ is now the standard deviation of the adjusted values of Y within the populations, and σ_m is the standard deviation of the adjusted population means: "The use and interpretation of η^2 as a proportion of variance and η as a correlation ratio now refers to Y', the dependent variable Y freed from that portion of its variance linearly associated with the covariate" (p. 373). It was mentioned in Section 2.3 that Tabachnick and Fidell (1989) and Winer et al. (1991) had made similar proposals for the application of the classical version of η^2 to analyses of covariance.

3.4. Estimating partial eta squared

Although this discussion was couched in terms of population parameters, Cohen's (1965) earlier account of η^2 had referred solely to sample statistics. In this case, Formula (6) would seem to constitute the appropriate generalisation of the sample statistic η^2 to complex factorial designs. More generally, Cohen (1965) maintained that any value of F with df1 and df2 degrees of freedom that had been obtained from sample data could be expressed as a correlation ratio using the following formula (see also Friedman, 1968):

$$\eta = \sqrt{\left[\frac{(\mathrm{df1} \cdot F)}{(\mathrm{df1} \cdot F + \mathrm{df2})}\right]}. \tag{7}$$

Equivalently (Cohen, 1973),

$$\eta^2 = \frac{(\mathrm{df1} \cdot F)}{(\mathrm{df1} \cdot F + \mathrm{df2})}. \tag{8}$$

In Section 2.2, I mentioned the idea that the correlation ratio in a sample would tend to overestimate the corresponding population value. Cohen (1965) similarly suggested that, if his version of η were calculated from sample data, it would tend to overestimate the relevant population value. Instead, he presented a variant of Kelley's (1935) ε which he claimed was an unbiased estimate of the latter. However, like ε, this was based on estimating the value of a fraction by means of inserting unbiased estimates of its numerator and denominator, and so it would have underestimated the relevant population value

Table 2
Simulated effects of an additional factor on effect size estimates.

Source of variation	SS	df	MS	F	p	η^2	Partial η^2
One-way analysis with Factor A alone							
Factor *A*	25.00	1	25.00	32.67	.000	.25	.25
Within groups	75.00	98	0.76				
Total	100.00	99					
Two-way analysis with Factor B that reduces the within-groups variation							
Factor *A*	25.00	1	25.00	48.00	.000	.25	.33
Factor *B*	25.00	1	25.00	48.00	.000	.25	.33
A × *B* interaction	0.00	1	0.00	0.00	n.s.	.00	.00
Within groups	50.00	96	0.52				
Total	100.00	99					
Two-way analysis with Factor B that introduces between-groups variation							
Factor *A*	25.00	1	25.00	32.00	.000	.20	.25
Factor *B*	25.00	1	25.00	32.00	.000	.20	.25
A × *B* interaction	0.00	1	0.00	0.00	n.s.	.00	.00
Within groups	75.00	96	0.77				
Total	125.00	99					

Adapted from).
SS, sum of squares; MS, mean squares; n.s., not significant.

of η. Cohen did not subsequently mention the idea that his version of η might be a biased estimate of the population value. Indeed, the illustrative examples that he provided in his 1969 book indicate that he intended the above account to apply equally to the computation of statistics from sample data. That account went through two later editions of the book essentially unchanged. Evidently, Cohen saw no need to qualify or amend it in the light of subsequent discussion.

Cohen (1973) stressed that he was not proposing his version of η^2 as an alternative to the classical version: rather, they were different measures of effect size with different goals. He described a version of Formula (1) as "quite properly the formula for η^2", but he described Formula (8) as "a formula for *partial* η^2" (p. 108). Henceforth, I refer to the latter as "partial η^2". Formulae (4) and (6) are simply special cases of this for one-way and factorial designs. Indeed, Formula (8) is equally appropriate both for research designs involving one or more covariates and for within-subjects designs (cf. Olejnik & Algina, 2000). Finally, it should be noted that Cohen's (1969) benchmarks for small, medium and large effects were intended to apply to partial η^2, not to the classical version, a point to be discussed later in this paper.

4. Metric properties of eta squared and partial eta squared

By definition, both the classical version of η^2 as defined in Formula (1) and partial η^2 as defined in Formula (6) vary between 0 and 1. The classical version of η^2 takes the value of 0 when the independent variable explains none of the variance in the dependent variable, and it takes the value of 1 when the independent variable in question explains *all* of the variance in the dependent variable. In a factorial design, the values of η^2 corresponding to the different components of variation are additive and nonoverlapping. Indeed, if one includes a value of η^2 for the variation within the groups (i.e., SS[Within groups]/SS[Total]), then the values of η^2 for the different components will add up to 1 (Kennedy, 1970; Levine & Hullett, 2002). Consequently, it makes sense to interpret values of η^2 as percentages of the total variance, as I did earlier in Section 2.1.

Partial η^2 does not share these properties. First, note that in a factorial design

$$[\mathrm{SS}(A) + \mathrm{SS}(\text{Within groups})] \leq \mathrm{SS}(\text{Total})$$

with equality only when none of the other independent variables and interactions explains any of the variance in the dependent variable. (This is, of course, the case in a single-factor design, where there are no other independent variables or interactions.) It follows that for a particular component of variance partial $\eta^2 \geq$ classical η^2. Partial η^2 takes the value of 1 when SS(Within groups) = 0 (in other words, when all of the independent variables and interactions in the design explain all of the variance in the dependent variable). Moreover, the values of partial η^2 for the different components of variance may add up to a value greater than 1, even if one does not include a value for the variance within the groups. (Examples of this found in published research are discussed in Section 5 below.) This makes it less sensible to interpret values of partial η^2 as percentages of the total variance.

Kennedy (1970) argued that in practice the value of partial η^2 depended on the nature and number of additional variables included in the research design, and that consequently it was difficult to make meaningful comparisons between the results obtained using different designs. Cohen (1973) acknowledged that it would limit the usefulness of partial η^2 if the additional variables served to explain (and therefore reduce) the variance within the groups. Table 2 shows three examples taken from an article by Levine and Hullett (2002). The top part of the table shows the results of a hypothetical one-way study for which η^2 and partial η^2 are both equal to .25. The middle part of the table shows the results of a hypothetical replication of this study in which a second factor *B* is controlled that had previously contributed to the within-groups variance. For these data, η^2 still equals .25, but partial η^2 equals .33, despite the fact that it is measuring exactly the same effect as in the first study.

Cohen (1973) suggested that it was appropriate to use η^2 only in cases such as this where "later experiments remove variance due to sources which had operated in earlier experiments by introducing them explicitly as control factors" (p. 111). However, he argued that in practice changes to the research design would serve instead to incorporate variables that had previously been held constant. These might be treatment variables that could be manipulated by the researcher or classification variables that could be investigated by the suitable selection of participants. They would be likely to introduce additional variation without affecting the within-groups variance. The bottom part of Table 2 shows the results of a hypothetical replication of the original study in which a second factor B is controlled that had not previously contributed to the within-groups variance; for these data, partial η^2 still equals .25, but η^2 is now only .20. Cohen concluded that partial η^2 should be used when additional manipulated or control variables were incorporated into the research design.

Sechrest and Yeaton (1982) went further than Kennedy and argued that partial η^2 could not be used even to compare the effects of different factors within the *same* design because they had different bases or denominators, an argument also put forward by Olejnik and Algina (2000). However, this is overstating the point. If $SS(A) > SS(B)$, it follows that

$$\frac{SS(A)}{SS(\text{Total})} > \frac{SS(B)}{SS(\text{Total})},$$

and so η^2 is larger for Factor A than for Factor (B). Yet it also follows that

$$SS(A) \cdot SS(B) + SS(A) \cdot SS(\text{Within Groups}) > SS(A) \cdot SS(B) + SS(B) \cdot SS(\text{Within Groups}).$$

Hence

$$SS(A) \cdot [SS(B) + SS(\text{Within groups})] > SS(B) \cdot [SS(A) + SS(\text{Within groups})] \quad \text{and}$$

$$\frac{SS(A)}{[SS(A) + SS(\text{Within groups})]} > \frac{SS(B)}{[SS(B) + SS(\text{Within groups})]}.$$

Thus, partial η^2, too, is larger for Factor A than for Factor B. At an ordinal level, at least, both η^2 and partial η^2 can be used to compare the effects of different factors in the same design.

Nevertheless, this argument cannot be made when comparing the effects of within-subjects variables. If A and B are independent variables in a study with a repeated-measures design, then partial η^2 takes the form $SS(A)/[(SS(A) + SS(A \times Ss)]$ and $SS(B)/[(SS(B) + SS(B \times Ss)]$. In this situation, the bases or denominators really are different, and so the two values of partial η^2 are incommensurable. The same is true in a mixed (split-plot) design: if A is the between-subjects variable and B is the within-subjects variable in such a design, then partial η^2 takes the form $SS(A)/[SS(A) + SS(\text{Within groups})]$ and $SS(B)/[(SS(B) + SS(B \times Ss)]$ (see Olejnik & Algina, 2000). In short, partial η^2 cannot be used to compare the effects of within-subjects variables either with one another or with the effects of between-subjects variables. To address this problem, Olejnik and Algina (2003) proposed a new measure of effect size that they called "generalised eta squared": this is akin to η^2 (Formula (1)) but variation due to classification variables (i.e., variables that are measured rather than manipulated) is removed from SS(Total) before determining the proportion of variance explained. This was endorsed by Bakeman (2005) but has yet to find favour among researchers. An alternative solution, of course, would be to use classical η^2: for both repeated-measures designs and mixed designs, this takes the form $SS(A)/SS(\text{Total})$ and $SS(B)/SS(\text{Total})$, which clearly can be compared.

5. Confusion between eta squared and partial eta squared

For many years, partial η^2 was not widely discussed in statistics textbooks. Levine and Hullett (2002) examined more than 20 such books that had been published between 1970 and 2000. They found that only one mentioned partial η^2 by name (Pedhazur, 1997, p. 507), but two others presented Formula (7) (that is, partial η) as a definition of classical η and Formula (6) (that is, partial η^2) as a definition of classical η^2 (Rosenthal & Rosnow, 1985, p. 14; 1991, p. 352). Levine and Hullett concluded that "there seems to be much confusion in the literature regarding eta squared and partial eta squared" (p. 613).

Further confusion surrounds the benchmarks suggested by Cohen (1969, pp. 278–280) to define small, medium, and large effects. As was explained earlier, these were based upon values of f that correspond to values of partial η^2 of .0099, .0588, and .1379, respectively. Nowadays, researchers often quote the latter values without explaining their derivation from values of f, which suggests a spurious degree of precision and makes them sound even more arbitrary than Cohen had originally intended. They are quoted as benchmark values for both partial η^2 and classical η^2, despite the fact that the former is typically greater than the latter. Indeed, Olejnik and Algina (2000) claimed that Cohen (1988, pp. 280–287) had suggested values of .01, .06, and .14 to indicate small, medium, or large effects for any measure of the proportion of variance explained. This is doubly inaccurate: first, because the figures are rounded to two decimal places (presumably to reduce the impression of spurious precision); and, second, because Cohen was clearly referring only to partial η^2 in offering such figures. (He provided different benchmarks for other measures of proportion of explained variance.)

As mentioned earlier, the values of partial η^2 for the different components of variance in a research design may add up to a value greater than 1, whereas values of classical η^2 will not. It follows that any examples of empirical research where the sizes of different effects add up to more than 1 must be based on partial η^2, even if they purport to be based

on classical η^2. Sechrest and Yeaton (1982) suggested that in practice observed effect sizes were typically of small absolute magnitude, and so actual examples of this would be scarce. However, Levine and Hullett (2002) found one study in the field of communication research where merely the values of η^2 that were reported added up to 3.98. They also used Formula (8) to show that the values of η^2 reported in three other studies in the same field were in fact values of partial η^2.

Similarly, Pierce, Block, and Aguinis (2004) inspected the articles published in "premier psychology journals" in 2001 and 2002. They found eight studies in six different articles where the values of η^2 reported in each study added up to more than 1. In one case, enough information was provided to reconstruct the complete analysis of variance summary table, and so Pierce et al. were able to confirm that the true values of classical η^2 were far smaller than those reported. They concluded that in these studies the purported values of classical η^2 were actually values of partial η^2. They concurred with Levine and Hullett that the misreporting of partial η^2 as classical η^2 was probably common but that in most cases it was likely to go unnoticed.

Levine and Hullett (2002) identified a possible cause of this confusion in a reporting error in the statistical package SPSS. Early versions of SPSS (such as SPSS^X and SPSS/PC+) did not yield values of η^2 or partial η^2 as optional output in their factorial analysis procedures. In SPSS Version 6 for Windows, the general factorial analysis procedure included an option for computing effect-size measures, and the output correctly labelled these as values of partial η^2 (Norušis & SPSS Inc., 1994, pp. 39, 41). However, Levine and Hullett found that factorial analyses of variance carried out with SPSS Version 9 for Windows yielded measures of effect size that were labelled "eta squared" but were actually the corresponding values of partial η^2. Moreover, the online help menu for the relevant option gave advice that confounded the two measures of effect size:

> *Estimates of effect size* gives a partial eta-squared value for each effect and each parameter estimate. The eta-squared statistic describes the proportion of the total variability attributable to a factor.

As Levine and Hullett pointed out, this explanation provided no definition of "partial eta-squared" (although the latter was correctly defined in the help menu for the procedure's PRINT options). As a consequence, even researchers who had read the online documentation would be likely to misreport values of partial η^2 as values of classical η^2 when using SPSS to analyse the data obtained using factorial designs. This problem seems to have arisen with the introduction of a new general linear model procedure for factorial analyses of variance with SPSS Version 7 for Windows. It affected versions of SPSS for Windows up to and including Version 10 (Pierce et al., 2004), but the labelling of the output was corrected in Version 11, released in 2001 (see Brace, Kemp, & Snelgar, 2003, pp. 159, 161). Even so, the help menu for the *Estimates of effect size* option continued to offer the same confusing advice.

This confusion affected textbooks on SPSS used by educational researchers and their students. For instance, in the first edition of *Discovering Statistics Using SPSS for Windows*, Field (2000, p. 299) stated that the general linear model procedure in SPSS Versions 7, 8 and 9 for Windows produced values of (classical) η^2. This error was carried over into the second edition of his book, which was concerned with SPSS Version 13 for Windows (Field, 2005, pp. 369, 384, 417), despite the fact that the relevant measures of effect size were by this point being explicitly labelled as values of partial eta squared. It was however corrected in the third edition of the book, concerned with SPSS Version 17 for Windows, which explained that the general linear model procedure yielded values of partial η^2 (Field, 2009, pp. 415–416).

Pierce et al. (2004) suggested that the problem was not limited to researchers who used SPSS, because the authors of one of the articles that they cited reported that they had used SAS to analyse their data. In fact, the general linear model procedure in SAS does not provide values of either η^2 or partial η^2. It is actually quite simple for a researcher to write a few lines of command syntax to generate these values, but which measure is generated and how it is labelled in the SAS output depend entirely on the researcher's own understanding and appreciation of the distinction between η^2 and partial η^2, not on the system itself.

Did the mislabelling of SPSS output cause similar problems in educational research? As a representative sample of studies, I examined those published in the year 2000 in Volume 10 of *Learning and Instruction*, the sister journal of *Educational Research Review*. Out of 26 articles, 11 reported inferential statistics that might have yielded measures of effect size (t or F). One presented effect sizes that reflected the percentages of total variance accounted for and are unambiguously values of classical η^2 (Blöte, Klein, & Beishuizen, 2000). Another used the standardised mean difference as a measure of effect size rather than a measure of explained variance in comparing older and younger adolescents (Bornholt, 2000). None of the nine remaining articles provided measures of effect size. In short, authors contributing to *Learning and Instruction* during 2000 were not confused by the mislabelling of SPSS output simply because it was not common practice at that time to report measures of effect size.

I also examined studies published in Volume 19 of *Learning and Instruction* in 2009 and found a completely different picture. Out of 43 articles, 25 contained inferential statistics that might have been associated with measures of effect size, and all 25 articles provided such measures. This is surprising in itself, because at the time of writing *Learning and Instruction* refers prospective authors to the American Psychological Association's *Publication Manual* but has no explicit policy with regard to the reporting of measures of effect size. Moreover, in each of the 25 articles, the reported measures of effect size were explicitly labelled "partial η^2". In short, within a period of less than 10 years, partial η^2 has gone from being barely used as a measure of effect size among quantitative researchers who contributed to *Learning and Instruction* to a position of virtually total hegemony. Doubtless this has been helped by the ready availability (and correct labelling) of partial η^2 in more recent versions of SPSS.

6. Limitations of eta squared and partial eta squared

A number of commentators have pointed out limitations of both η^2 and partial η^2 as measures of effect size (O'Grady, 1982; Pedhazur, 1997, pp. 505–509; Sechrest & Yeaton, 1982). In Section 4, I noted that both η^2 and partial η^2 have a logical upper bound of 1. To take the hypothetical example in Table 1, the only situation in which η^2 and partial η^2 could equal 1 is one in which all the boys obtain one score and all the girls obtain some other score. If the dependent variable is distributed in any other way, then η^2 and partial η^2 will have an upper bound of less than 1. For instance, if the dependent variable is normally distributed, it can be shown that both η^2 and partial η^2 have an upper limit of approximately .64 (Pedhazur, 1997, p. 507). In addition, if the dependent variable is not perfectly reliable, measurement error will contribute to the within-group variability, and this will reduce still further the proportion of variation that can in principle be explained (Sechrest & Yeaton, 1982).

η^2 and partial η^2 also depend on the choice and number of levels of the independent variable. Fisher (1925, p. 219) pointed out that, when the latter is theoretically continuous, the value of η^2 obtained from a particular sample would depend not only upon the range of values that is explored but also upon the number of values employed within that range. More generally, Lindquist (1953) argued that the differentiation ratio "depends upon the arbitrary choice of categories in the treatment classifications, and hence is not meaningful as an index of strength of relationship" (p. 63). Levin (1967) noted in particular that the percentage of explained variance could be artificially inflated by the inclusion of a treatment group that was known to produce a substantially different level of performance. O'Grady (1982) suggested that, as a general rule, the more diverse a population is in terms of the factor of interest, the higher will be the estimates of explained variance in the dependent variable.

Sechrest and Yeaton (1982) suggested that measures of explained variance would be sensitive to the properties of the participants, materials and procedure used in each study:

> As a general proposition it can be stated that *all measures of variance accounted for are specific to characteristics of the experiments from which the estimates were obtained*, and therefore the ultimate interpretation of proportion of variance accounted for is a dubious prospect at best. (p. 592, italics in original)

Consequently, attempted replications might yield different results because of the uncontrolled sources of variation contributing to relevant error terms.

Pedhazur (1997, p. 506) agreed with this suggestion and also argued that comparisons among effects within the same study would not be meaningful unless the manipulations of different variables could be shown to be of comparable magnitude. Suppose, for instance, that reading ability is studied in 8-year-old and 12-year-old boys and girls, with the result that $\eta^2 = .40$ for gender and $\eta^2 = .20$ for age. This could simply mean that the manipulation of gender was stronger than the manipulation of age. In practice, however, researchers usually have no way of assessing the strength of their manipulations. Indeed, it may not even be meaningful to compare two strengths of manipulation: what range of ages *would* represent an equivalent manipulation to the comparison of boys and girls (cf. Grissom & Kim, 2005, p. 142–143)?

Classical η^2 and partial η^2 are both sample statistics and are subject to measurement error. Strictly speaking, therefore, any reported values should be accompanied by confidence intervals (cf. American Educational Research Association, 2006; Wilkinson & Task Force on Statistical Inference, 1999). In fact, confidence limits can be calculated for the noncentral F distribution, and these could then be used to derive confidence intervals for both classical η^2 and partial η^2 (Fowler, 1985; Kline, 2004, pp. 118–121; Smithson, 2001, 2003, chap. 4).

7. Conclusions

Recent commentators have arrived at different positions with regard to the usefulness of classical η^2 and partial η^2. Field (2009, pp. 415–416) and Kline (2009, pp. 166–168) just presented the two measures of effect size without explaining why one might be preferred to the other. Indeed, Kline envisaged that researchers might report values of *both* η^2 *and* partial η^2 in their research publications, but this is surely likely to confuse rather than enlighten their readers. Levine and Hullett (2002) were influenced by the reporting error exacerbated by the mislabelling of SPSS output and by the argument put forward by Kennedy (1970) against the use of partial η^2. They suggested that "researchers should most often report eta squared, omega squared, or epsilon squared rather than partial eta squared" (p. 623). The opposite point of view was adopted by Bakeman (2006), who advocated the exclusive use of partial η^2 except when comparing effects of between-subjects and within-subjects variables. A more conciliatory approach was adopted by Olejnik and Algina (2000) and by Pierce et al. (2004), who were content for researchers to use either classical η^2 or partial η^2 provided that they were both explicit in their choice of measure and accurate in their calculations.

Table 3 summarises the various characteristics of classical η^2 and partial η^2 that have been discussed in this review. Neither measure demonstrates an unequivocal advantage over the other, although some considerations favour partial η^2: that it can be applied to all research designs; that it can be calculated from inferential statistics in published research reports; that it can be benchmarked against Cohen's (1969) suggested criteria of small, medium and large effects; and that it is readily available in SPSS. These considerations are probably sufficient to explain and justify the position of hegemony enjoyed by partial η^2 in educational research.

Table 3
Characteristics of classical eta squared and partial eta squared.

Classical eta squared	Partial eta squared
Can be readily calculated from results of a factorial analysis of variance: SS(*A*)/SS(Total).	Can be readily calculated from results of a factorial analysis of variance: SS(*A*)/[SS(*A*) + SS(Within groups)].
Does not generalise readily to designs involving covariates or within-subjects designs.	Generalises to all designs.
Cannot be calculated from inferential statistics in published reports except for one-way between-subjects designs.	Can be readily calculated from inferential statistics in published reports: (df1·*F*)/(df1·*F* + df2).
In a factorial design, the values are additive and nonoverlapping and cannot sum to more than 1.	In a factorial design, the values may sum to more than 1.
Is unaffected by the inclusion of variables that explain the within-groups variation.	Is increased by the inclusion of variables that explain the within-groups variation.
Is reduced by the inclusion of variables that introduce additional variation.	Is unaffected by the inclusion of variables that introduce additional variation.
Can be used to compare the effects of different between-subjects or within-subjects factors in the same design.	Can only be used to compare the effects of different between-subjects factors in the same design.
No benchmarks have been suggested, although Cohen's (1969, pp. 278–280) criteria are sometimes incorrectly cited.	Can be benchmarked against Cohen's (1969, pp. 278–280) criteria of small, medium and large effects.
Not readily available in SPSS except for one-way between-subjects designs.	Readily available in SPSS.
Not widely used in educational research.	Widely used in educational research.

How appropriate are Cohen's criteria? Analogous benchmarks that he suggested for other measures of effect size have been criticised for being too strict on the basis of what one might reasonably expect to find in actual research (Hemphill, 2003; Lipsey & Wilson, 1993). However, Cooper and Findley (1982) calculated values of Cohen's *f* for 185 research studies cited in textbooks of social psychology and found an average value of 0.51, which is larger than Cohen's (1969) benchmark of 0.40 for "large" effects. Haase, Waechter, and Solomon (1982) calculated values of partial η^2 from 11,044 inferential statistics reported in the *Journal of Counseling Psychology* from 1970 to 1979 and found a median value of .0830, somewhat larger than Cohen's (1969) benchmark of .0588 for "medium" effects. For two different areas of research, then, there is no evidence that Cohen's benchmarks for partial η^2 are too strict.

Even so, this conclusion needs to be accompanied by certain health warnings. Some have suggested that in certain circumstances nonstatistical criteria may be more important for judging the theoretical or practical significance of empirical findings. Prentice and Miller (1992) suggested that measures of effect size were entirely appropriate when there was a broad consensus about the operationalisation of the independent variable and the choice of the dependent variable. Nevertheless, when there was more latitude over the research design, they argued that small effects might well be more pertinent (see also Abelson, 1985; Yeaton & Sechrest, 1981). On the other hand, Jacobson and Truax (1991) suggested that judgements about the efficacy of clinical treatments tended to be made by consumers, physicians and other stakeholders on the basis of external criteria that were far more stringent than purely statistical benchmarks.

Reporting classical η^2 or partial η^2 may be a way of satisfying overzealous editors or reviewers, but the interpretation of these measures needs to be undertaken with care. They can be influenced by the research design and by the selection of participants, and there are circumstances where it is inappropriate to make comparisons among the values obtained in different studies or even among the values obtained for different effects within the same study. In other words, calculating effect sizes is the beginning of the story, not the end.

Acknowledgements

I am grateful to Laurel Fisher, Paul Ginns, James Hartley, Anesa Hosein, Natalia Kucirkova, Erik Meyer, Anna Parpala, Richard Remedios and Dan Richards for their comments on previous versions of this paper.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129–133. doi:10.1037/0033-2909.97.1.129
American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33–40. doi:10.3102/0013189X035006033
Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379–384.
Bakeman, R. (2006). The practical importance of findings. *Monographs of the Society for Research in Child Development*, *71*(3), 127–145, doi:10.1111/j.1540-5834.2006.00408.x.
Blöte, A. W., Klein, A. S., & Beishuizen, M. (2000). Mental computation and conceptual understanding. *Learning and Instruction*, *10*, 221–247, doi:10.1016/S0959-4752(99)00028-6.

Bornholt, L. J. (2000). Social and personal aspects of self knowledge: A balance of individuality and belonging. *Learning and Instruction*, *10*, 415–429, doi:10.1016/S0959-4752(00)00006-2.
Brace, N., Kemp, R., & Snelgar, R. (2003). *SPSS for psychologists: A guide to data analysis using SPSS for Windows (Versions 9, 10 and 11)* (2nd ed.). Basingstoke, UK: Palgrave Macmillan.
Cohen, J. (1962). The statistical power of abnormal–social psychological research. *Journal of Abnormal and Social Psychology*, *65*, 145–153, doi:10.1037/h0045186.
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. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.
Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement*, *33*, 107–112. doi:10.1177/001316447303300111
Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). New York: Academic Press.
Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin*, *8*, 168–173. doi:10.1177/0146167282810 26
Diamond, S. (1959). *Information and error: An introduction to statistical analysis*. New York: Basic Books.
Field, A. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.
Field, A. (2005). *Discovering statistics using SPSS (and sex, drugs and rock 'n' roll)* (2nd ed.). London: Sage.
Field, A. (2009). *Discovering statistics using SPSS (and sex and drugs and rock 'n' roll)* (3rd ed.). London: Sage.
Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
Fowler, R. L. (1985). Point estimates and confidence intervals in measures of association. *Psychological Bulletin*, *98*, 160–165. doi:10.1037/0033-2909.98.1.160
Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, *70*, 245–251, doi:10.1037/h0026258.
Glass, G. V., & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Educational Research Journal*, *6*, 403–414. Retrieved from http://www.jstor.org/stable/1161859
Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, *29*, 58–65, doi:10.1037/0022-0167.29.1.58.
Halperin, M., Hartley, H. O., & Hoel, P. G. (1965). Recommended standards for statistical symbols and notation. *American Statistician*, *19*(3), 12–14. Retrieved from http://www.jstor.org/stable/2681417
Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart, & Winston.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*, 78–79, doi:10.1037/0003-066X.58.1.78.
Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Belmont, CA: Thomson Wadsworth.
Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*, 227–240, doi:10.1177/0013164402062002002.
Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19, doi:10.1037/0022-006X.59.1.12.
Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, *21*, 554–559. Retrieved from http://www.jstor.org/stable/86523
Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics: Vol. 1. Distribution theory* (4th ed.). London: Charles Griffin.
Kennedy, J. J. (1970). The eta coefficient in complex ANOVA designs. *Educational and Psychological Measurement*, *30*, 885–889, doi:10.1177/001316447003000409.
Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioural research*. Washington, DC: American Psychological Association.
Kline, R. B. (2009). *Becoming a behavioural researcher: A guide to producing research that matters*. Washington, DC: American Psychological Association.
Levin, J. R. (1967). Misinterpreting the significance of "explained variation". *American Psychologist*, *22*, 675–676, doi:10.1037/h0037681.
Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*, 612–625, doi:10.1111/j.1468-2958.2002.tb00828.x.
Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education*. Boston, MA: Houghton Mifflin.
Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209, doi:10.1037/0003-066X.48.12.1181.
McNemar, Q. (1962). *Psychological statistics* (3rd ed.). New York: Wiley.
Norušis, M., & SPSS Inc. (1994). *SPSS advanced statistics 6.1*. Chicago, IL: SPSS.
O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, *92*, 766–777. doi:10.1037/0033-2909.92.3.766
Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286. doi:10.1006/ceps.2000.1040
Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*, 434–447. doi:10.1037/1082-989X.8.4.434
Pearson, K. (1903). Mathematical contributions to the theory of evolution: On homotyposis in homologous but differentiated organs. *Proceedings of the Royal Society of London*, *71*, 288–313. Retrieved from http://www.jstor.org/stable/116386
Pearson, K. (1905). *Mathematical contributions to the theory of evolution: XIV. On the general theory of skew correlation and non-linear regression (Drapers' Company Research Memoirs, Biometric Series, No. II)*. London: Dulau.
Pearson, K. (1923). On the correction necessary for the correlation ratio, η. *Biometrika*, *14*, 412–417, doi:10.1093/biomet/14.3-4.412.
Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.
Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, *64*, 916–924, doi:10.1177/0013164404264848.
Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, *112*, 160–164, doi:10.1037/0033-2909.112.1.160.
Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments, and Computers*, *28*, 12–22.
Richardson, J. T. E. (2007). Measuring the relationship between scores on two questionnaires. *Educational Research Review*, *2*, 13–27, doi:10.1016/j.edurev.2006.08.002.
Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge, UK: Cambridge University Press.
Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.
Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. *Evaluation Review*, *6*, 579–600, doi:10.1177/0193841X8200600501.
Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632, doi:10.1177/00131640121971392.
Smithson, M. J. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.
Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: HarperCollins.
Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, *54*, 837–847.
Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*(2), 26–30, doi:10.3102/0013189X025002026.
Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance. *Educational Psychology Review*, *11*, 157–169, doi:10.1023/A:1022028509820.

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education*, *70*, 80–93.
Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604, doi:10.1037/0003-066X.54.8.594.
Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.
Winkler, R. L., & Hays, W. L. (1975). *Statistics: Probability, inference, and decision* (2nd ed.). New York: Holt, Rinehart, & Winston.
Yeaton, W. H., & Sechrest, L. (1981). Meaningful measures of effect. *Journal of Consulting and Clinical Psychology*, *49*, 766–767, doi:10.1037/0022-006X.49.5.766.