Effect Size Estimates: Current Use, Calculations, and Interpretation

Article *in* Journal of Experimental Psychology General · August 2011

DOI: 10.1037/a0024338 · Source: PubMed

3 authors, including:

Catherine Fritz
The University of Northampton
26 PUBLICATIONS **1,088** CITATIONS

SEE PROFILE

Journal of Experimental Psychology: General
2012, Vol. 141, No. 1, 2–18

0096-3445/11/$12.00 DOI: 10.1037/a0024338

Effect Size Estimates: Current Use, Calculations, and Interpretation

Catherine O. Fritz and Peter E. Morris
Lancaster University

Jennifer J. Richler
Vanderbilt University

The *Publication Manual of the American Psychological Association* (American Psychological Association, 2001, 2010) calls for the reporting of effect sizes and their confidence intervals. Estimates of effect size are useful for determining the practical or theoretical importance of an effect, the relative contributions of factors, and the power of an analysis. We surveyed articles published in 2009 and 2010 in the *Journal of Experimental Psychology: General*, noting the statistical analyses reported and the associated reporting of effect size estimates. Effect sizes were reported for fewer than half of the analyses; no article reported a confidence interval for an effect size. The most often reported analysis was analysis of variance, and almost half of these reports were not accompanied by effect sizes. Partial η^2 was the most commonly reported effect size estimate for analysis of variance. For *t* tests, 2/3 of the articles did not report an associated effect size estimate; Cohen's *d* was the most often reported. We provide a straightforward guide to understanding, selecting, calculating, and interpreting effect sizes for many types of data and to methods for calculating effect size confidence intervals and power analysis.

Keywords: effect size, eta squared, confidence intervals, statistical reporting, statistical interpretation

Experimental psychologists are accomplished at designing and analyzing factorial experiments and at reporting inferential statistics that identify significant effects. In addition to statistical significance, most research reports describe the direction of an effect, but it is also instructive to consider its size. Estimates of effect size are useful for determining the practical or theoretical importance of an effect, the relative contribution of different factors or the same factor in different circumstances, and the power of an analysis. This article reports the use of effect size estimates in the 2009 and 2010 volumes of the *Journal of Experimental Psychology: General* (*JEP: General*), comments briefly on their use, and offers practical advice on choosing, calculating, and reporting effect size estimates and their confidence intervals (CIs).

Effect size estimates have a long and somewhat interesting history (for details, see Huberty, 2002), but the current attention to them stems from Cohen's work (e.g., Cohen, 1962, 1988, 1994) championing the reporting of effect sizes. In response to Cohen (1994) the American Psychological Association (APA) Board of Scientific Affairs set up a task force that proposed guidelines for statistical methods for psychology journals (Wilkinson & the APA Task Force on Statistical Inference, 1999). These guidelines were subsequently incorporated into the revised fifth edition of the *Publication Manual of the American Psychological Association* (APA, 2001; hereinafter *APA Publication Manual*) and were again included in the sixth edition (APA, 2010). Regarding effect sizes, the sixth edition states,

> For the reader to appreciate the magnitude or importance of a study's findings, it is almost always necessary to include some measure of effect size in the Results section. Whenever possible, provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size. (APA, 2010, p. 34)

Effect sizes allow researchers to move away from the simple identification of statistical significance and toward a more generally interpretable, quantitative description of the size of an effect. They provide a description of the size of observed effects that is independent of the possibly misleading influences of sample size. Studies with different sample sizes but the same basic descriptive characteristics (e.g., distributions, means, standard deviations, CIs) will differ in their statistical significance values but not in their effect size estimates. Effect sizes describe the observed effects; effects that are large but nonsignificant may suggest further research with greater power, whereas effects that are trivially small but nevertheless significant because of large sample sizes can warn researchers against possibly overvaluing the observed effect.[1] Effect sizes can also allow the comparison of effects in a single study and across studies in either formal or informal meta-analyses. When planning new research, previously observed effect sizes can be used to calculate power and thereby estimate appropriate sample sizes. Cohen (1988), Keppel and Wickens (2004), and most statistical textbooks provide guidance on calculating power; a very brief, elementary guide appears in the Appendix along with mention of planning sample sizes based on accuracy in

This article was published Online First August 8, 2011.

Catherine O. Fritz, Educational Research Department, Lancaster University, Lancaster, United Kingdom; Peter E. Morris, Department of Psychology, Lancaster University, Lancaster, United Kingdom; Jennifer J. Richler, Department of Psychology, Vanderbilt University.

We thank Thomas D. Wickens and Geoffrey Cumming for their very helpful advice on an earlier version of this article.

Correspondence concerning this article should be addressed to Catherine O. Fritz, Educational Research Department, Lancaster University, Lancaster LA1 4YD, United Kingdom. E-mail: c.fritz@lancaster.ac.uk

[1] It is rarely the case that experimental studies have the problem of too many cases making trivial effects statistically significant, but some large-scale surveys and other studies with very large sample sizes can have this problem. For example, a correlation of .1, accounting for only 1% of the variability, is statistically significant with a sample size of 272 (one tailed).

parameter estimation (i.e., planning the size of the CIs; Cumming, 2012; Kelley & Rausch, 2006; Maxwell, Kelley, & Raush, 2008).

A brief note on the terminology used in this article may be helpful. Effect sizes calculated to describe the data in a sample, like any other descriptive statistic, also potentially estimate the corresponding population parameter. Throughout this article, we refer to the calculated effect size, which describes the sample and estimates the population, as an *effect size estimate*. It is important to remember that the estimates both describe the sample and estimate the population and that some statistics, as we describe later, provide better estimation of the population parameters than do others.

The most basic and obvious estimate of effect size when considering whether two data sets differ is the difference between the means; most articles report means, and the difference is easily calculated. Some researchers argue that differences between the means are generally sufficient and superior to other ways of quantifying effect size (e.g., Baguley, 2009; Wilkinson & the APA Task Force on Statistical Inference, 1999). The raw difference between the means can provide a useful estimate of effect size when the measures involved are meaningful ones, such as IQ or reading age, assessed by a standard scale that is widely used. Discussion of the effect would naturally focus on the raw difference, and it would be easy to compare the results with other research using the same measure.

However, comparing means without considering the distributions from which the means were calculated can be seriously misleading. If two studies (A and B) each have two conditions with means of 100 and 108, it would be very misleading to conclude that the effects in the two studies are the same. If the standard deviations for the conditions in Study A were both two and in Study B were both 80, then it is clear that the distributions for Study A would have virtually no overlap, whereas those for Study B would overlap substantially. Using Cohen's U_1, which we describe later, we find that only 2% of the distributions for Study A would overlap, given the standardized difference between the means (d) of 4, but 92% of the distributions would overlap in Study B because the standardized difference between these means is $d = 0.1$. Significance tests make the difference between the two studies quite clear. For the study with standard deviations of two, a t test would find a two-tailed significant difference ($p < .05$) with three participants per group, but 770 participants per group would be needed to obtain a significant difference for the study with standard deviations of 80. The consequence of the difference in the size of the distributions is also obvious when considering the CIs: With 50 samples in each study, Study A's CI = ± 0.6, whereas Study B's CI = ± 22.2. These examples illustrate how comparisons between means without considering the variability of the data can conceal important properties of the effect. To address this problem, standardized effect size calculations have been developed that consider variability as well as the differences between means. Effect size calculations are addressed by a growing number of specialized texts, including Cumming (2012), Ellis (2010), Grissom and Kim (2005, 2011), and Rosenthal, Rosnow, and Rubin (2000) as well as many general statistical texts.

When examining the difference between two conditions, effect sizes based on standardized differences between the means are commonly recommended. These include Cohen's d, Hedges's g, and Glass's d and Δ. When independent variables have more than two levels or are continuous, effect size estimates usually describe the proportion of variability accounted for by each independent variable; they include eta squared (η^2, sometimes called R^2), partial eta squared (η_p^2), generalized eta squared (ω_G^2), associated omega squared measures (ω^2, ω_p^2, ω_G^2), and common correlational measures, such as r^2, R^2, and $R_{adj}{}^2$. In addition, there are other less frequently encountered statistics, such as epsilon squared (ε^2; Ezekiel, 1930) and various statistics devised by Cohen (1988), including q, f, and f^2. Finally, there are the effect size estimates relevant to categorical data, such as phi (ϕ), Cramér's V (or ϕ_c), Goodman–Kruskal's lambda, and Cohen's w (Cohen, 1988). The plethora of possible effect size estimates may create confusion and contribute to the lack of engagement with reporting and interpreting effect sizes. Many of these statistics are conceptually, and even algebraically, quite similar but have been developed as improvements or to serve different types of data and different purposes. The emergence of a consensus to use a few selected estimates would probably be a useful simplification, as long as the choice was driven by the genuine usefulness of those estimates and not merely by their easy availability.

One important distinction to make among effect sizes is that some statistics, such as η^2 and R^2, describe the samples observed but may overestimate the population parameters, whereas others, such as ω^2 and adjusted R^2, attempt to estimate the variability in the sampled population and, thus, in replications of the experiment. These latter statistics are often recommended by statistical textbooks because they relate to the population and are less vulnerable to inflation from chance factors. However, researchers very rarely report these population estimates, perhaps because they tend to be smaller than the sample statistics.

Although the *APA Publication Manual* has strongly advocated the reporting of effect sizes for 10 years and many psychology editors have done so for longer than that (e.g., Campbell, 1982; Levant, 1992; Murphy, 1997), a glance through many journals suggests that such reporting is inconsistent. Morris and Fritz (2011) surveyed cognitive articles published in 2009; they found that only two in five of these articles intentionally reported effect sizes. Isabel Gautier, as the incoming editor of *JEP: General*, asked us to conduct a similar survey of recent volumes of this journal and to review the methods of calculating effect size estimates.

Method

We reviewed articles published in the 2009 and 2010 volumes of *JEP: General*, noting the statistical analyses, descriptive statistics, and effect size estimates reported in each.

Results

Table 1 provides frequencies of the most commonly used statistical analyses for each year; corresponding percentages are illustrated in Figure 1. Note that data are reported for each article, not for each experiment, but the analyses were similar across experiments in most articles. Analysis of variance (ANOVA) was reported in most articles, 83% overall, followed by t tests, 66% overall; these were often used together to locate the source of effects in factorial designs. Overall, 58% of the articles reported at least one measure of effect size (73% for 2009, 45% for 2010).

Table 1
Numbers of Articles Reporting Commonly Used Statistical Analyses

Year	Articles	ANOVA	*t* test	Correlation	Regression
2009	33	27	23	14	8
2010	38	32	24	13	9
Overall	71	59	47	27	17

Note. Corresponding percentages are shown in Figure 1. ANOVA = analysis of variance.

Different analyses are often associated with different estimates of effect size; therefore, the use of specific effect size estimates is reported within the context of the analysis conducted. Table 2 shows the effect size estimates used in conjunction with ANOVA type analyses, including analysis of covariance. Overall, slightly more than half of the articles reported a measure of effect size associated with ANOVA at least once; η_p^2 was by far the most frequently used, almost certainly because it is provided by SPSS. Effect size estimates were rarely reported for further ANOVA-related analyses, such as simple effects and post hoc and planned comparisons. Table 3 summarizes the frequency of further ANOVA-related analyses and the inclusion of effect size estimates. Most articles did not report all components of the ANOVA; only 10 of the 59 articles (five from each year) reported the mean square error (*MSE*) terms and only 20 (12 for 2009 and 8 for 2010) reported *F* ratios for all effects. When reporting ANOVA, articles usually included descriptive statistics for the data, either in terms of individual cells or marginals. At least some means associated with ANOVA were reported in 93% of the articles (93% for 2009 and 94% for 2010); some measure of variability was less often reported, appearing in only 80% of the articles (81% for 2009 and 78% for 2010).

When reporting *t* tests, roughly one quarter of articles included a measure of effect size; Cohen's *d* was the most often used effect size estimate. See Table 4 for numbers and percentages of effect size estimates and descriptive statistics reported in association with *t* tests. Descriptive statistics were less often provided than for ANOVA; almost one quarter did not report a measure of central tendency, and almost half failed to report the variability of the data.

Neither intentional reporting of effect size estimates nor descriptive statistics tended to accompany reports of correlations. Refer to Table 5 for numbers and percentages of articles intentionally reporting effect size estimates and descriptive statistics associated with correlation analyses. Fewer than 10% of the articles reporting correlations provided r^2 or any other associated effect size estimate beyond the correlation, and fewer than one quarter reported descriptive statistics for the data. Although *r* is a useful estimate of effect size, there is a difference between reporting it as a correlation and treating it as an estimate of effect size; none of these articles appeared to present it as an effect size estimate.

Various types of regression were also reported in 17 articles. Most of these were very selective in terms of the statistics reported from the analysis and in terms of descriptive statistics; there were almost no reports of effect size. Table 6 shows numbers and percentages of statistics reported in association with regression analyses. Most articles did not report the *F* ratio or significance value for the test, although most reported some statistics associated with the predictors, such as the *t* tests, the regression weights, the partial correlations, or the odds ratios. We counted all reports of R^2 as estimates of effect size, although most were not explicitly presented as such.

A few nonparametric and frequency-based tests were also reported; only one of these included a measure of effect size. These reports also tended to neglect statistical summaries of the data.

Discussion

Our initial concern over the reporting of effect sizes was justified by our analysis. Across the 2 years studied, 42% of articles reported no measure of effect size. Most of the articles counted as including effect sizes reported them for only some of the analyzed effects. Even where articles reported η_p^2 for ANOVA analyses, they often omitted effect size estimates for nonsignificant effects and other comparisons. Fewer than a third of the articles reporting *t* tests included associated effect size estimates to aid in interpreting the results. On the positive side, reported effect sizes in the *JEP: General* articles were clear with respect to which effect size statistic was used. Our recent survey of cognitive articles (Morris & Fritz, 2011) found articles in which effect sizes were wrongly identified, and we have occasionally encountered articles that report effect size figures without identifying which statistic was used. As Vacha-Haase and Thompson (2004) observed, it is essential to correctly identify which statistic is used: Reporting that an effect size is .5 means something very different depending on whether the statistic used is *d*, *r*, r^2, η^2, ω^2, or others.

We observed almost no interpretation of the effect sizes that were reported, despite APA's direction to address the "theoretical, clinical, or practical significance of the outcomes" (APA, 2010, p. 36). Clearly effect sizes are important in a clinical and practical sense. Are they less relevant in a theoretical sense? If theories are solely concerned with the statistical significance of effects and not with their size, then perhaps there is no useful role for effect size consideration in interpretation, but surely good theories are concerned with substantive significance rather than merely statistical significance. A theory that only predicts a difference (or relation-

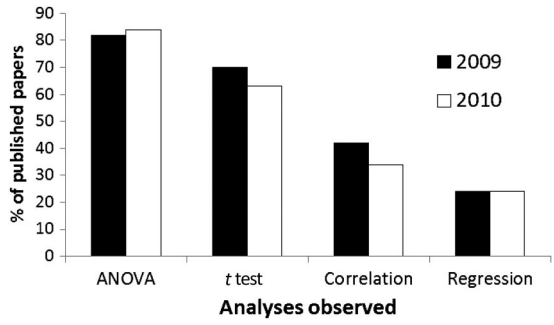

Figure 1. Percentage of published articles including each type of analysis. Other types of analyses were also observed with lower frequencies, including χ^2 (18% for 2009 and 16% for 2010), nonparametric difference tests (3% for 2009 and 11% for 2010), and Cronbach's alpha (6% for 2009 and 11% for 2010); these were not accompanied by measures of effect size and are not discussed here. Corresponding frequencies are shown in Table 1. ANOVA = analysis of variance.

Table 2
Number (and Percentage) of Articles Reporting Effect Size Estimates Associated With ANOVA

Year	Articles with ANOVA	Any ES measure	η^2	η_p^2	ω^2	ω_p^2	*d*
2009	27	18 (67)	1 (6)	17 (94)	0	0	2 (11)
2010	32	15 (47)	5 (33)	9 (60)	1 (7)	0	3 (20)
Overall	59	33 (56)	6 (18)	26 (79)	1 (3)	0	5 (15)

Note. Articles were included in the counts if the statistic was reported at least once. Percentages across measures sum to more than 100% because some articles included more than one measure of effect size. ANOVA = analysis of variance; ES = effect size.

ship) but is not concerned with the size of that effect will be one that is quite difficult to falsify and perhaps even more difficult to apply.

It appears that effect sizes may be reported to meet the minimum letter of the law, with little regard for the spirit of the law. The preponderance of η_p^2 in these analyses and the sparsity of discussion of reported effect sizes is consistent with a scenario wherein people obtain η_p^2 values from their statistical software and report it as required, but they give scant consideration to the implications of the values obtained. Little appears to have changed in the 60 years since Yates (1951) observed that an emphasis on statistical significance testing had two main negative consequences: Statisticians develop significance tests for

> problems . . . of little or no practical importance [and] scientific research workers . . . pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating. (p. 32).

Many researchers may be cautious about engaging too deeply with the effect size values that they calculate because, in contrast to the use of inferential statistics, they have far less experience in using the effect size estimates as an aspect of evaluating results and providing guidance for future research. In part, this situation may arise from the tendency to report η_p^2, which has limited usefulness. The η_p^2 statistic may be useful for cross-study comparisons with identical designs, but where designs differ, η_G^2 is needed. Within a factorial study, η_p^2 cannot properly be used to compare effects; η^2 is needed. We describe each of these measures and discuss the possible uses and interpretations of various effect sizes in a later section of the article.

The fifth and sixth editions of the *APA Publication Manual* recommend the inclusion of CIs for effect size estimates, but none were included in any of the *JEP: General* articles that we examined, and these CIs were only reported in 1 of the 386 cognitive articles that we surveyed (Morris & Fritz, 2011). Effect sizes, like means, are point estimates. They describe the sample and provide an estimate of the population parameter. For an estimate to be useful, it is important to provide some idea of how precise that estimate might be—the expected range within which the population parameter falls with some specified probability. When means are reported, it is widely accepted good practice to report some measure of variability—either the standard error, the standard deviation (from which the standard error is easily calculated), or a CI. These variability statistics provide a guide to the probable values for the population parameter. The effect size estimate also requires some accompanying description of its likely variability; that variability statistic is the associated CI.

The lack of these CIs in research reports is, however, understandable. Textbooks that describe effect size statistics typically do not provide associated guidance for calculating the CIs. Commonly used statistical software packages also fail to provide them. Furthermore, most measures of effect size are noncentrally distributed (see, e.g., Ellis, 2010, pp. 19–21; Grissom & Kim, 2005, p. 64), a somewhat nonintuitive concept that makes them more difficult to understand and to calculate. The CIs above and below the effect size estimate are not equal in size and have to be estimated by special software carrying out iterative procedures (e.g., Cumming, 2012; Cumming & Finch, 2001; Smithson, 2003; Steiger, 2004). These unusual characteristics of CIs for effect size estimates, combined with researchers' lack of familiarity with them, may help to explain why these CIs are not reported. In a later section on CIs, we suggest sources for relevant software and offer formulas to approximate CIs for Cohen's *d* and R^2.

Analyses were sometimes reported without the relevant descriptive statistics, making it more difficult for the reader to understand and evaluate the results. The most basic sort of effect size estimate when evaluating differences is the difference between means, but almost one quarter of *t* test reports were not accompanied by means, and almost half lacked reports of variability measures. When evaluating correlations, it is also necessary to consider the

Table 3
Articles Reporting Additional Analyses Associated With ANOVA and Effect Size Reporting for Those Analyses

Year	Articles with ANOVA	Post hoc or planned contrasts	Post hoc or planned effect size	For significant interactions		
				No. of articles	Simple effects	Simple effect size
2009	27	16	1 (6%)	18	10 (55%)	1 (10%)
2010	32	20	1 (5%)	23	14 (44%)	2 (14%)
Overall	59	36	2 (6%)	41	24 (41%)	3 (13%)

Note. Articles were included in the counts if the analysis or statistic appeared at least once. ANOVA = analysis of variance.

Table 4
Number (and Percentage) of Articles Reporting Effect Size Estimates and Descriptive Statistics Associated With t Tests

Year	Articles with *t* test	Effect size estimates		Descriptive statistics	
		d	η_p^2	*M*	Variability
2009	23	9 (39)	0	18 (78)	16 (70)[a]
2010	24	3 (13)	2 (8)	18 (75)	10 (42)[b]
Overall	47	12 (26)	2 (4)	36 (77)	26 (55)

Note. Articles were included if the statistic was reported at least once. [a] Nine articles reported standard deviation, six reported standard error of the mean, and one reported 95% confidence interval. [b] Nine articles reported standard deviation, and five reported standard error of the mean.

distribution of the data, but the great majority—about 80%—of the correlation reports failed to include a description of the data.

The frequent use of ANOVA, reported in 83% of the *JEP: General* articles, resembled our finding for cognitive journals (Morris & Fritz, 2011) where 86% of articles reported at least one ANOVA. There is an argument that ANOVA is overused and that in many cases, regression would be more appropriate (e.g., Cohen, Cohen, West, & Aiken, 2003). ANOVA is a valuable technique, but when factors are categorized versions of continuous variables, it is less appropriate than more flexible techniques, such as multiple regression. After all, ANOVA and regression both derive from the general linear model, and ANOVA can be regarded as a special case of regression (e.g., Tabachnick & Fidell, 2007, p. 155). To fit the requirements of ANOVA, researchers sometimes treat continuous variables, such as age, skill level, and knowledge, as if there were a small number of discrete, categorical values. This process is problematic in that it loses some of the power inherent in the original continuous variable. The fewer the number of levels defined, the more data and thus the more power are lost. If just two or three levels are defined, it can also be tempting when planning factorial research to select groups or conditions that are high and low on some variable and, as a result, to almost certainly overestimate the influence of the variable and the size of its effect in the population including midrange values. In addition, although ANOVA tests the statistical significance of each individual effect,

Table 5
Number (and Percentage) of Articles Reporting Effect Size Estimates and Descriptive Statistics Associated With Correlations

Year	Articles with correlation	Effect size (r^2)	Descriptive statistics	
			M	Variability
2009	14	1 (7)	2 (14)	2 (14)[a]
2010	13	1 (8)	5 (38)	4 (31)[b]
Overall	27	2 (7)	7 (26)	6 (22)

Note. Articles were included if the statistic was reported at least once. Although *r* is also a measure of effect size, it was not described or treated as such in the articles surveyed.
[a] The articles reported standard deviation. [b] Two articles reported standard deviation and two reported standard error of the mean.

Table 6
Number (and Percentage) of Articles Reporting Effect Size Estimates and Descriptive Statistics Associated With Regression Analyses

Year	Articles with regression	R^{2a}	*F*	Descriptive statistics	
				M	Variability
2009	8	4 (7)	1 (13)	6 (75)	6 (75)
2010	9	2 (8)	3 (33)	3 (33)	2 (22)
Overall	17	2 (7)	4 (24)	9 (53)	8 (47)

Note. Articles were included if the statistic was reported at least once. [a] One of the 2009 articles reported adjusted R^2 as well as R^2; no other articles reported adjusted R^2.

multiple regression allows the effect of variables to be evaluated both collectively and in terms of their individual contributions. Statistical outputs of regression analyses typically include clear measures of effect size at all levels and in various forms, through R^2, adjusted R^2, changes in R^2, standardized regression weights, partial correlations and semipartial correlations. However, among the articles surveyed, regression was rarely used except as a model-fitting tool.

A substantial number of articles (20%) failed to report measures of the variability of data analyzed by ANOVA. As we illustrated earlier, standard deviations are as important as means for understanding data sets. One might argue that both standard deviations and standard errors should be reported, which rarely occurred. Where both were reported, it was usually the case that standard errors appeared as error bars on a figure and standard deviations were reported in the text or a table; this combination provided an accessible, clear description of the data. Like significance tests and CIs, the size of standard errors depends on the sample size, so that although a standard error is very valuable for interpreting differences between conditions, it does not provide an easily appreciated idea of the distribution of the data. Standard deviations are a more straightforward way of helping the reader conceptualize data sets.

Although the reporting of *MSE* is not a requirement for APA journals, many editors encourage it, so it was surprising that 83% of the articles surveyed did not include *MSE*s with the reports of ANOVAs. The great advantage of reporting *MSE*s is that, along with the value of the *F* ratio and its accompanying degrees of freedom, knowledge of the relevant *MSE* allows the reader to reconstruct the remaining ANOVA details for that test including the sums of squares, which are useful for effect size estimations. Thus, the reporting of *MSE*s opens up considerable opportunities for readers wishing to understand the data in greater depth and perhaps calculate their own effect size estimates. Ideally, for each ANOVA it would be valuable for the article to include the significance tests of all effects—both significant and nonsignificant—with their *MSE*s reported so that the full details of the analysis could be reconstructed. Complete reporting would be useful for meta-analyses and would allow readers to calculate the types of η^2 and ω^2 that they thought most appropriate.

Calculating Effect Sizes

Our aim in this section is to demystify as far as possible selected effect size estimates and to recommend convenient ways of cal-

culating them. General purpose statistics books for psychologists (e.g., Aron, Aron, & Coups, 2009; Howell, 2002) typically address few effect size statistics and may not consider them thoroughly. Readers may prefer to consult specialized statistics books addressing effect sizes (e.g., Cumming, 2012; Ellis, 2010; Grissom & Kim, 2005; Rosenthal et al., 2000).

This article addresses several effect sizes: those specific to comparing two conditions (Cohen's *d*, Hedges's *g*, Glass's *d* or Δ, and point biserial correlation *r*), those describing the proportion of variability explained (η^2, η_p^2, η_G^2, R^2, the ω^2 family, adjusted R^2, and ε^2), and effect sizes for nonnormal data (*z* associated with the Mann–Whitney and Wilcoxon tests, and ϕ, Cramér's *V*, and Goodman–Kruskal's lambda for categorical data).

Effect Sizes Specific to Comparing Two Conditions

The most common approach to calculating effect size when comparing two conditions is to describe the standardized difference between the means, that is, the difference between the means of the two conditions in terms of standard (*z*) scores. There are varieties of this approach, discussed later, based on the way the standard deviation is calculated. In all cases, the sign of the effect size statistic is a function of the order assigned to the two conditions; where the conditions are not inherently ordered, a positive effect size should be reported. Online calculators for the standardized difference statistics are available (e.g., Becker, 2000; Ellis, 2009).

Cohen's *d* and Hedges's *g*. Cohen (1962, 1988) introduced a measure similar to a *z* score in which one of the means from the two distributions is subtracted from the other and the result is divided by the population standard deviation (σ) for the variables:

$$d = \frac{M_A - M_B}{\sigma},$$

where M_A and M_B are the two means and σ refers to the standard deviation for the population. Hedges (1982) proposed a small modification for his statistic *g* in which the population standard deviation (σ, calculated with a denominator of *n*, the number of cases) is replaced by the pooled sample standard deviation (*s*, calculated with a denominator of $n - 1$)

$$g = \frac{M_A - M_B}{s}.$$

The standard deviations made available by common statistical packages are for the sample(s) so that the more convenient statistic for researchers to calculate is *g* rather than *d*. However, as we observed in our review, it is rare for authors to report Hedges's *g*, even though it may be what they have actually calculated. It appears to be the case that *d* may be often used as a generic term for this type of effect size. For example, Borenstein, Hedges, Higgins, and Rothstein (2009) referred to the *g* statistic defined above as *d* as does Comprehensive Meta-Analysis software that is widely used for meta-analysis. These sources use *g* to refer to an *unbiased* calculation, sometimes called $d_{unbiased}$ or d_{unb}, that is particularly useful for small sample sizes, where *d* tends to overestimate the population effect size. The formula to adjust *d*, from Borenstein et al. (p. 27), is

$$g \text{ or } d_{\text{unb}} = d\left(1 - \frac{3}{4df - 1}\right).$$

The correction is very small when the sample size is large (only 3% for $df = 25$) but is more substantial with a smaller sample size (8% for $df = 10$). This value is not the same as the original Hedges's *g* (1982), described earlier, although *g* might be used to refer to either; d_{unb} is a less ambiguous symbol, but in either case the formula should be provided for clarity.

A discussion of the rather confusing history of the chosen symbols for these statistics can be found in Ellis (2010). For most reasonably sized samples, the difference between Cohen's *d*, calculated using *n*, and Hedges's *g*, calculated using $n - 1$ degrees of freedom (*df*), will be very small. Especially when sample sizes are small, it is helpful for authors to clearly specify how the reported effect size estimates were calculated, regardless of what symbol is used, so that the reader can interpret them correctly and they might be useful for subsequent meta-analyses.

There is virtually always some difference between the standard deviations of the two distributions. When the standard deviations (s_A and s_B) and the sample sizes of the two distributions (*A* and *B*) are very similar, it may be sufficiently accurate when estimating the combined standard deviation (s_{AB}) to take the average of the two standard deviations:

$$s_{AB} = \frac{s_A + s_B}{2}.$$

When the standard deviations differ but the sample sizes for each group are very similar, then averaging the square of the standard deviations and taking the square root of the result is more accurate (Cohen, 1988, pp. 43–44; Keppel & Wickens, 2004, p. 160):

$$s_{AB} = \sqrt{\frac{s_A^2 + s_B^2}{2}}.$$

However, where the sample size and/or the standard deviation of the two distributions differ markedly it is usually recommended (e.g., Keppel & Wickens, 2004) that the sums of squares and the degrees of freedom for the two variables should be combined with the following formula (Keppel & Wickens, p. 160):

$$s_{AB} = \sqrt{\frac{SS_A + SS_B}{df_A + df_B}}.$$

That is, the sum of squares for the two variables *A* and *B* should be added together, as should the degrees of freedom for the variables. Then, the sum of the sums of squares is divided by the sum of the degrees of freedom, and the square root of the result taken. When not provided by the statistical package, the sum of squares for a variable can be easily calculated from the standard deviation as

$$SS = df \times s^2$$

or from the standard error of the mean (*SE*) as

$$SS = df \times SE^2 \times N.$$

If pairs of conditions are being compared from among several that have been evaluated by an ANOVA, rather than working out the standard deviation for each comparison, it is acceptable to

replace the combined standard deviation for the multiple comparisons by the square root of the *MSE* (Grissom & Kim, 2005):

$$s_{AB} \approx \sqrt{MSE}$$

In an attempt to help with the interpretation of *d*, Cohen (1988) suggested that *d* values of .8, .5, and .2 represented large, medium, and small effect sizes, respectively, perhaps more meaningfully described as obvious, subtle, and merely statistical. He recognized that what would be a large, medium, or small effect size would, in practice, depend on the particular area of study, and he recommended these values for use only when no better basis for estimating the effect size index was available. These designations clearly do not reflect practical importance or substantive significance, as those are judgments based on a more comprehensive consideration of the research.

Glass's *d* or Δ. An alternative to both Cohen's *d* and Hedges's *g* involves using the standard deviation for a control group rather than a standard deviation based on combining the groups. This approach is appropriate if the experimental manipulations are thought to have distorted the distribution in some way. This measure was proposed by Glass (1976) and is known as Glass's *d* or Δ.

Point biserial correlation, *r*. There are alternatives to using the standardized difference statistics as described earlier. Some (e.g., Rosenthal, Rosnow, & Rubin, 2000) have preferred the point biserial correlation coefficient, *r*, on the grounds that psychologists are already familiar with it. Furthermore, r^2 is equivalent to η^2 and other effect size estimates that describe the proportion of variability associated with an effect, described later. For two groups, the point biserial correlation, *r*, is calculated by coding group membership with numbers, for example, 1 and 2. The correlation between these codes and the scores for the two conditions give the value of point biserial *r*. It is also easy to calculate *r* if an independent samples *t* test has already been carried out because

$$r = \sqrt{\frac{t^2}{t^2 + df}}.$$

Just as the sign of the *t* statistic is an artifact of the order assigned to the conditions, so too is the sign of the effect size. Unless there is a meaningful order for the two conditions, the statistics should be reported as positive numbers. If there is a meaningful order and it was used for the *t* test, the sign of the *t* statistic should be applied to *r*.

Table 7 provides values of *r* corresponding to values of *d* when group sizes are similar. An excellent discussion of the relative benefits and limitations of *d* and point biserial *r* is provided by McGrath and Meyer (2006). For point biserial *r*, McGrath and Meyer suggested that values of .37, .24, and .10 represent large, medium and small—or obvious, subtle, and merely statistical—effect sizes, respectively. Formulas for converting between several effect size estimates, including *r*, are provided in Table 8.

Table 7
Associated Values of Cohen's d, r, r^2 (or η^2), PS, and U_1

d	*r*	r^2 or η^2	*PS*	U_1
0.0	.00	.000	50	0
0.1	.05	.002	53	8
0.2	.10	.010	56	15
0.3	.15	.022	58	21
0.4	.20	.038	61	27
0.5	.24	.059	64	33
0.6	.29	.083	66	38
0.7	.33	.11	69	43
0.8	.37	.14	71	47
0.9	.41	.17	74	52
1.0	.45	.20	76	55
1.1	.48	.23	78	59
1.2	.51	.27	80	62
1.3	.55	.30	82	65
1.4	.57	.33	84	68
1.5	.60	.36	86	71
1.6	.63	.39	87	73
1.7	.65	.42	89	75
1.8	.67	.45	90	77
1.9	.69	.47	91	79
2.0	.71	.50	92	81
2.2	.74	.55	94	84
2.4	.77	.59	96	87
2.6	.79	.63	97	89
2.8	.81	.66	98	91
3.0	.83	.69	98	93
3.2	.85	.72	99	94
3.4	.86	.74	99	95
3.6	.87	.76	99	96
3.8	.89	.78	100	97
4.0	.89	.80	100	98

Note. *PS* = probability of superiority. *PS* is the percentage of occasions when a randomly sampled member of the distribution with the higher mean will have a higher score than a randomly sampled member of the other distribution. U_1 = the percentage of nonoverlap between the two distributions. Data are from Grissom (1994) and Cohen (1988); they assume similar sample sizes.

Effect Sizes Describing the Proportion of Variability Explained

For pairs of conditions, it is also possible to apply proportion of variability statistics such as R^2 or η^2, in a manner similar to the squared point biserial correlation, r^2, described earlier. We turn next to these variability-based measures. Most of the variability-based effect size estimates involve comparing various combinations of sums of squares and means squares taken from ANOVA summary tables.[2] To illustrate the different measures we refer to Table 9, which reports an imaginary three-way between-subjects ANOVA.

Partial eta squared (η_p^2). The η_p^2 statistic is simply the ratio of the sum of squares for the particular variable under consideration divided by the total of that sum of squares and the sum of squares of the relevant error term. It describes the proportion of variability associated with an effect when the variability associated with all other effects identified in the analysis has been removed from consideration. As we described earlier, it is the most commonly reported effect size in recent issues of *JEP: General*. This popularity is almost certainly because η_p^2 can be calculated directly by SPSS. In general, the formula is

[2] Sums of squares are calculated by subtracting the mean for any set of data from each score, squaring each result, and summing these squared deviations from the mean. The mean square is the sum of squares divided by the degrees of freedom.

Table 8
Formulas for Deriving Effect Size Estimates Directly and Indirectly

From this statistic	To this statistic: *d*	To this statistic: Point biserial *r*	To this statistic: η^2 with similar group sizes
Direct formula	$d = \frac{M_A - M_B}{\sigma}$	$r = \sqrt{\frac{SS_{effect}}{SS_{total}}}$	$\eta^2 = \frac{SS_{factor}}{SS_{total}}$
d	—	$r = \frac{d}{\sqrt{d^2 + 4}}$	$\eta^2 = \frac{d^2}{d^2 + 4}$
Point biserial *r*	$d = \frac{2r}{\sqrt{1 - r^2}}$	—	$\eta^2 = r^2$
η^2 with similar group sizes	$d = \frac{2\sqrt{\eta^2}}{\sqrt{(1 - \eta^2)}}$	$r = \sqrt{\eta^2}$	—
t with similar group sizes	$d = \frac{2t}{\sqrt{N - 2}}$	$r = \sqrt{\frac{t^2}{t^2 + df}}$	$\eta^2 = \frac{t^2}{t^2 + df}$

Note. When group sizes differ considerably (when one group has fewer than one third of the total *N*), then *r* is smaller than the above calculation. For more information about the translation between statistics with very uneven sample sizes, see McGrath and Meyer (2006).

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}.$$

As an example, in Table 9, Factor A has a sum of squares of 100, and the error term has a sum of squares of 600, so

$$\eta_p^2 = \frac{100}{100 + 600} = .14.$$

Apart from asking SPSS to calculate η_p^2, it is easy to abstract the necessary sums of squares from ANOVA summary tables reported by statistical software. It is also easy to calculate η_p^2 from an *F* ratio and its degrees of freedom, because

$$\eta_p^2 = \frac{df_{effect} \times F_{effect}}{(df_{effect} \times F_{effect}) + df_{error}}.$$

Thus, it is possible to calculate η_p^2 from published results where the authors have not reported this effect size.

Care must be taken when comparing η_p^2 estimates across studies with different designs to ensure that the error terms are comparable. The size of η_p^2 is influenced by changes to the *error* variability. Error variability (SS_{error}) increases when sources of variability are neither controlled nor identified as part of the analysis; it decreases when these sources of variability are controlled or are identified in the analysis. The variability associated with an uncontrolled variable appears in the SS_{error}, thereby reducing the size of η_p^2, whereas controlling that variable or including it as an individual differences factor in the analysis removes that variability from the SS_{error}, thereby increasing the value of η_p^2. Some of these issues can be addressed by using η_G^2, which we discuss later.

One can use η_p^2 to compare the effect of some factor that appears in multiple studies but only when the error terms are comparable; η_G^2 (see later) is more generally useful for between-study comparisons. For comparing the relative contribution of different factors within a single study, η_p^2 is not useful because the baseline variability (i.e., the denominator) is different for each calculation.

Eta squared (η^2). The η^2 statistic (sometimes called R^2) is a simple ratio of the variability associated with an effect compared with all of the variability in an analysis.[3] It describes the proportion of the total variability in the data that are accounted for by the effect under consideration. One can easily calculate η^2 from the ANOVA output from a statistical package; it is the ratio of the sum of squares for the effect divided by the total sum of squares. That is,

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}.$$

Thus, in Table 9, η^2 for Factor A is calculated from the sum of squares for A (100) and the total sum of squares (1,280). So, for Factor A,

$$\eta^2 = \frac{100}{1280} = .08.$$

One can also calculate η^2 from reported *F* ratios and degrees of freedom where all of the effects from an ANOVA are reported. In a two-factor (G × H) between-groups design, for the effect of Factor G

$$\eta^2 = \frac{df_{effect_G} \times F_G}{(df_{effect_G} \times F_G) + (df_{effect_H} \times F_H) + (df_{effect_{G \times H}} \times F_{G \times H}) + df_{error}}.$$

[3] Some authors prefer to refer to η^2 as R^2 because it fits with the statistical convention of reserving Greek letters for population parameters and because of the commonality with R^2 for regression. Although it seems simpler to use just one term for the proportion of variability explained, statistical software and textbooks most often use η^2 for ANOVA and R^2 for regression. Furthermore, η_p^2, which is more often used than η^2, is equivalent to a partial correlation—a concept that is less familiar to people who do not use multiple regression regularly. We choose to use η^2 with ANOVA for these reasons and because there is an argument that applying one term to ANOVA and another to regression is clearer and simpler.

Table 9
Example Between-Groups Analysis of Variance Summary Table With Calculations of η^2, η_p^2, ω^2, and ω_p^2

Source	*SS*	*df*	*MS*	*F*	η^2	η_p^2	ω^2	ω_p^2
Factor A	100	1	100	9.3	.08	.14	.07	.12
Factor B	200	1	200	18.7	.16	.25	.15	.22
Factor C	50	1	50	4.7	.04	.08	.03	.06
A × B	100	1	100	9.3	.08	.14	.07	.12
A × C	20	1	20	1.9	.02	.03	.01	.01
B × C	10	1	10	0.9	.01	.02	0 (−.001)[a]	0 (−.002)[a]
A × B × C	200	1	200	18.7	.16	.25	.15	.22
Error	600	56	10.7					
Total	1,280	63						

Note. *SS* = sum of squares; *MS* = mean squares.
[a] Negative values of ω^2 can occur when $F < 1$. Keppel and Wickens (2004) recommend setting the value to zero.

The denominator term (SS_{total}) sums the degrees of freedom for the error term with the products of each *F* ratio and its corresponding degrees of freedom.

The η^2 statistic is a useful measure of the contribution of an effect—of a factor or an interaction—to the observed dependent variable. So, for example, examining the values of η^2 in Table 9 reveals that Factor B accounts for twice as large a proportion of the total variability as Factor A, but also that the three-way interaction of Factors A, B, and C contributes as much variability as does Factor B.

For comparing the size of effects within a study, η^2 is useful, but there are risks in comparing η^2 values across studies with different designs. These risks derive from the differences in total variability that arise from manipulating additional variables, thereby adding variability, or from controlling variables, thereby reducing variability. If the effect of Factor A is the same across two studies (i.e., SS_A remains constant), a study that manipulates Factor A alone will have a greater value for η^2 than one that manipulates Factor A and introduces an additional manipulated Factor B. This difference is because in the latter case, the total variability is increased by the variability introduced with Factor B. Conversely, controlling variables so that they do not contribute their variability to the overall ANOVA will, obviously, reduce the SS_{total}. If the controlled variables do not interact with an effect, so that the SS_{effect} is unchanged, then the η^2 for that effect will be larger than if the variables had not been controlled.

Unmatched total variability is an issue for cross-study comparisons involving most measures of effect size. Cohen's *d*, for example, depends on the standard deviations of the variables and they, in turn, depend on the extent to which other factors have been controlled.

Psychologists calculating η^2 for their own data for the first time are often disappointed by the size of the effect that they are studying. A manipulation with an η^2 of .04 accounts for only 4% of the total variability in the dependent variable—an amount that may seem trivial, especially when compared to r^2 values commonly seen in correlational research. It may be easier to deal with small η^2 values in terms of Cohen's (1988, pp. 283–287) description of large (.14), medium (.06), and small (.01) effects, but obviously it is the practical or theoretical importance of the effect that determines what size qualifies the outcome as substantively significant. In most experimental research, observed effect sizes are likely to be small; many factors influence behavior in almost any area, and few of these will be examined in the analysis. It would be an exceptional situation to research a behavior that was determined by only one or two causal factors. Each factor makes its own contribution to the total variability under consideration. If several factors vary together, they may jointly account for a substantial proportion of the variability, but any individual factor might contribute only a relatively small part of the whole. Alternative calculations, described later, produce variants of η^2 that eliminate some of the other variability from consideration.

Generalized eta squared (η_G^2). Scientific research is a cumulative activity; it is necessary to compare and combine the results of research across studies. Unfortunately, neither η^2 nor η_p^2 is well suited for making comparisons across studies with different designs. η_G^2 provides a way to compare effect size across studies; it was introduced by Olejnik and Algina (2003) and Bakeman (2005) extended their description of its use for repeated measures designs. Like R^2 and η^2, η_G^2 gives an estimate of the proportion of variability within a study that is associated with a variable but without the distorting effects of variables introduced in some studies but not others. For Olejnik and Algina, the distinction between manipulated factors and individual differences factors is key. To illustrate the distinction, a study that tested children in two different types of experimental rooms would have room type as a manipulated variable. However, if the children from a class were classified into groups by their ages and by their personalities, these would be individual differences factors. The central idea when calculating η_G^2 is that the sums of squares for manipulated variables are not included in the denominator of the calculation, except under two conditions. Those conditions are (a) when calculating η_G^2 for the manipulated variable itself and (b) when calculating η_G^2 for an interaction between that manipulated variable and either an individual differences factor or a subject factor in a repeated measures design (i.e., the between-subjects error term).

We can demonstrate the calculation of η_G^2 using the Table 9 example. Suppose that, continuing our developmental example, Factor A is the room type in which the children are tested, Factor B is age group (younger or older children from the class), with two levels, and Factor C is a two-level classification of the children, such as introvert or extravert. Factor A is a manipulated factor but Factors B and C are individual differences factors. To calculate η_G^2 for Factor B (Age, an individual differences factor), use the formula for η^2 but remove from the total sums of squares in the denominator the sums of squares associated with Factor A because it is a manipulated factor—one that adds variability to the design. Thus, although η^2 for Factor B is

$$\eta^2 = \frac{SS_B}{SS_{Total}} = \frac{200}{1280} = .16,$$

the adjusted calculation for η^2_G is

$$\eta^2_G = \frac{SS_B}{SS_{Total} - SS_A} = \frac{200}{1280 - 100} = \frac{200}{1180} = .17.$$

The adjustment, removing the variability that was added by the manipulated variable, results in a higher value for η^2_G; more importantly, it is a value that can be compared with the η^2_G value from another study, even if the designs of the two studies differed.

As a further example, suppose that Factor C was, instead, a manipulated variable such as presence or absence of an adult in the room. In this case, the η^2_G for Factor B will remove the variability associated with both manipulated variables (A and C) and their interaction (A × C):

$$\eta^2_G = \frac{SS_B}{SS_{Total} - SS_A - SS_c - SS_{AxC}} = \frac{200}{1280 - 100 - 50 - 20} = \frac{200}{1110} = .18.$$

Similar calculations can be easily made for repeated measures designs, although the denominator may have to be constructed by accumulating the appropriate sums of squares rather than by subtraction from the total sum of squares. These sums of squares can be obtained as part of the analysis from the statistical software or, for published work, by reconstructing the ANOVA summary table (such as in Table 9) if the reporting of the ANOVA was sufficiently complete—that is, if all effects were reported complete with all mean squares for the error terms (*MSE*s).

As an example of constructing η^2_G for a repeated measures factor, imagine that an analysis involves just two repeated measures factors, P and Q, with both as manipulated factors. When creating the denominator for the effect size of Factor P, the sum of squares for Q will be omitted. However, the denominator will include all sources of variability associated with P or with the between-subjects error: the sums of squares for P, for subjects (Subj), and for the interactions P × Subj, Q × Subj, and P ×Q × Subj. So, for P,

$$\eta^2_G = \frac{SS_P}{SS_P + SS_{Subj} + SS_{P\times Subj} + SS_{Q\times Subj} + SS_{P\times Q\times Subj}}.$$

Details of the appropriate sums of squares to be included in the denominator for most common designs can be found in Bakeman (2005).

***R*².** In regression, R^2 is the square of the correlation between the observed values and the values predicted by the regression equation. It is used to report the proportion of the variability of the dependent variable that is predictable from the set of variables entered into the regression and thus provides a good effect size estimate. R^2 is calculated from the ratio

$$R^2 = \frac{SS_{Regression}}{SS_{Total}}.$$

R^2 is similar to η^2 in that the variability associated with the focus of the analysis—in this case the prediction—is considered as a proportion of the total variability; R^2 and η^2 are identical when the predictor is a factor, coded as a dummy variable. Changes in R^2 when new variables are added in hierarchical regressions allow the contributions of independent variables to be assessed.

$$R^2_{Change} = \frac{SS_{Change}}{SS_{Total}}.$$

The square of the semipartial ("part" in SPSS) correlation between an independent variable and the dependent variable when the other independent variables have been controlled gives the proportion of the total variability uniquely predicted by the independent variable—analogous to η^2. Similarly, the square of the partial correlation between an independent variable and the dependent variable is analogous to η^2_p. Thus, in multiple regression analyses, R^2, R^2_{Change}, the squared semipartial correlations and the squared partial correlations answer many questions about the size of the relative contributions of the dependent variables. Note that although R^2 and η^2 are the same in that each describes the proportion of the total variability that is accounted for, their use is somewhat different. In regression, R^2 describes the effect of a *set* of variables (one or more), whereas in ANOVA, η^2 describes the effect of a *single* factor or interaction (equivalent to the squared semipartial correlation in regression).

ω^2, ω^2_p, and ω^2_G. The various η^2 and R^2 statistics describe the effect size observed in the research. However, it is often valuable to think beyond a particular study, to the population from which the sample came, and therefore of the effect size that would be predicted in a replication of the study. In a replication, the variability accounted for by each factor or set of predictors is likely to be somewhat different from the observations from one sample. Sample variability includes both population variability and sampling variability and so tends to be somewhat larger than the population value alone. Thus, R^2 overstates the variation in the population, especially for small effects. Various statistics have therefore been developed to estimate the effect size in the population rather than the observed sample.

One statistic that is popular with the authors of statistical textbooks (e.g., Hays, 1973; Howell, 2002; Keppel & Wickens, 2004; Tabachnick & Fidell, 2007) is ω^2. However, our survey of recent *JEP: General* articles and of articles from three 2009 cognitive journals (Morris & Fritz, 2011) found that despite the recommendations of these and other textbook authors, it is very rare for an article to report ω^2. We observed only one instance in the combined set of 457 articles.

As for η^2, there are three types of ω^2 estimates: ω^2, ω^2_p, and ω^2_G. The basic principle of ω^2 is that it is the ratio of the population variability explained by the factor being measured to the population's total variability. For a one-way ANOVA, the total variability can be divided into the variability associated with a particular factor and the error variability. So, for a one way ANOVA with Factor A,

$$\omega^2 = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_{error}},$$

where σ^2_A represents the population variance for Factor A, and σ^2_{error} represents the appropriate population error variance for Factor A. The same formula applies when calculating ω^2_p, where the error term is the term against which the relevant factor is evaluated.

The basic formula for estimating ω^2 in a one-way ANOVA or ω_p^2 in a factorial design is

$$\omega^2 \text{ or } \omega_p^2 \text{ for } A = \frac{SS_{effect} - (a-1) \times MS_{error}}{SS_{total} + MS_{error}},$$

where a is the number of levels of the factor (Hays, 1973, p. 513). This same value can be calculated directly from the F statistic (Keppel & Wickens, 2004, p. 233):

$$\omega^2 \text{ or } \omega_p^2 \text{ for } A = \frac{(a-1) \times (F_A - 1)}{(a-1) \times (F_A - 1) + N},$$

where a is the number of levels of the Factor A, and N is the total number of subjects.

One can calculate ω^2 in a similar way for multifactor between-subject designs. The numerator remains the same, but the denominator includes the product of the degrees of freedom and the F ratio reduced by 1 for each of the effects (factors and interactions) in the analysis. So, for a multifactor design,

$$\omega^2 = \frac{(a-1) \times (F_{effect} - 1)}{\Sigma[df_{effect} \times (F_{effect} - 1)] + N},$$

summing across all of the effects (Keppel & Wickens, 2004, p. 481).

We have used the formulas to calculate ω^2 and ω_p^2 for each of the factors in Table 9. For these particular imaginary data, ω^2 and η^2 are similar and so are ω_p^2 and η_p^2. This near identity is because the example has a reasonable sample size, with just two levels for each factor, and the effect itself is large. The size of the distortion for sample rather than population effect size calculations (i.e., η^2 rather than ω^2) depends on the number of participants tested, the number of levels of the factors, and the size of the effect. More participants, fewer levels, and larger effects lead to less difference between ω^2 and η^2. With reasonably sized samples, limited numbers of factor levels, and larger effects, the overestimation of η^2 may often be acceptable. This is fortunate, because there are problems in estimating ω^2 for repeated measures designs; for these, only a range, not the actual value, can be calculated (Keppel & Wickens, 2004, p. 427). Instead, η^2 has to be reported, but the inflation of the estimate has to be recognized. Advice on calculating ω_G^2 can be found in Olejnik and Algina (2003).

R^2_{adj} and ε^2. For the R^2 calculated by multiple regression, there has long been the Wherry (1931) formula for calculating adjusted or shrunken R^2 (R^2_{adj}) with the aim of predicting, like ω^2, the R^2 to be expected if the study were to be repeated with a sample from the same population.

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{N-1}{N-k-1}\right),$$

where N is the sample size, and k is the number of independent variables in the analysis. Many statistical software packages calculate R^2_{adj}.

A similar approach is taken to calculating an effect size known as ε^2 (Ezekiel, 1930), which is an alternative to ω^2. However, ε^2 is rarely reported, and we do not discuss it further here. Details of its calculation can be found in Richardson (1996).

Effect Sizes for Nonparametric Data

Effect size estimates for Mann–Whitney and Wilcoxon nonparametric tests. Most of the effect size estimates we have described here assume that the data have a normal distribution. However, some data do not meet the requirements of parametric tests, for example, data on an ordinal but not interval scale. For such data, researchers usually turn to nonparametric statistical tests, such as the Mann–Whitney and the Wilcoxon tests. The significance of these tests is usually evaluated through the approximation of the distributions of the test statistics to the z distribution when sample sizes are not too small, and statistical packages, such as SPSS, that run these tests report the appropriate z value in addition to the values for U or T; z can also be calculated by hand (e.g., Siegel & Castellan, 1988). The z value can be used to calculate an effect size, such as the r proposed by Cohen (1988); Cohen's guidelines for r are that a large effect is .5, a medium effect is .3, and a small effect is .1 (Coolican, 2009, p. 395). It is easy to calculate r, r^2, or η^2 from these z values because

$$r = \frac{z}{\sqrt{N}},$$

and

$$r^2 \quad or \quad \eta^2 = \frac{z^2}{N}.$$

These effect size estimates remain independent of sample size despite the presence of N in the formulas. This is because z is sensitive to sample size; dividing by a function of N removes the effect of sample size from the resultant effect size estimate.

Effect sizes for categorical data. Categorical data are often tested with the chi-square statistic (χ^2) but, like ANOVA and t tests, the significance of a χ^2 test depends on the sample size as well as the strength of the association. There are various measures of association for contingency tables; we describe three that may be used for unordered categories. These can be easily calculated using SPSS by choosing Analyse, Descriptive Statistics, Crosstabs, Statistics and choosing the appropriate statistic.

Where the data being analyzed are in a 2×2 contingency table the ϕ correlation coefficient can be used. One can calculate ϕ from χ^2 for the data using the formula

$$\phi = \sqrt{\frac{\chi^2}{N}},$$

where N is the total sample size. If, for example, the obtained value of χ^2 was 10 with a sample size of 40 then

$$\phi = \sqrt{\frac{10}{40}} = \sqrt{0.25} = .05.$$

Cramér (1946) extended the ϕ statistic to larger contingency tables than the 2×2 of the ϕ correlation. This statistic, known as Cramér's V or ϕ_c, modifies the formula for ϕ to be

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}},$$

where N is the total sample size, and k is the number of rows or columns in the table, whichever is the smaller. Be aware that, unlike Pearson's r, the square of ϕ, or of Cramér's V, is not a valid description of the proportion of variability accounted for (Siegel & Castellan, 1988, p. 231).

When the rows and columns of a contingency table represent a predictor and a predicted variable, Goodman–Kruskal's lambda (L) describes how much the prediction is improved by knowing the category for the predictor, a potentially useful description of the size of the effect (Ellis, 2010; Siegel & Castellan, 1988). One may calculate lambda from any size contingency table; two values can be calculated: how well the row variable improves the predictability of the column variable and vice versa. Usually only one direction is meaningful. To calculate L_{row} for predicting column membership from row membership, sum the highest frequency in each of the columns, subtract the largest row total, and divide by the total number of observations not in that largest row. The formula is

$$L_{row} = \frac{\Sigma_{j=1}^{k} n_{Mj} - max(R_i)}{N - max(R_i)},$$

where k is the number of columns, n_{Mj} is the highest frequency in the jth column, $\max(R_i)$ is the largest row total, and N is the total number of observations (Siegel & Castellan, 1988, p. 299). So, for example, to determine how much attending a seminar improved the ability to predict an adequate answer on the relevant exam question, the contingency table might appear as in Table 10. One would calculate lambda as

$$L_{row} = \frac{(75 + 30) - 90}{144 - 90} = \frac{15}{54} = .28,$$

so knowing row membership improves prediction of answer quality by 28%. Notice that lambda can also be zero as in the altered data in Table 11. Here, lambda is calculated as

$$L_{row} = \frac{(75 + 15) - 90}{144 - 90} = \frac{0}{54} = 0.0.$$

Where knowledge of the row does not contribute to predicting column membership, lambda is zero. The lambda statistic seems especially useful in describing the size of the effect in terms that people without statistical training are likely to easily understand.

CIs for Effect Sizes

As discussed earlier, the calculation of CIs for effect sizes is not as straightforward as it is for means because the distributions are not centered on the effect size value. Help is available, though. Cumming (2012) provided guidance and Excel-based software for calculating CIs for *d,* which can be downloaded from http://www.thenewstatistics.com. Bird (2002) described methods for calculating effect sizes for ANOVA, and Smithson (2003) provided instructions and downloadable scripts for SPSS, SAS, SPlus, and R for calculating CIs for effect sizes associated with t tests, ANOVA, regression, and χ^2 analyses at http://dl.dropbox.com/u/1857674/CIstuff/CI.html. This webpage also provides links to other websites that may be helpful. These calculators include consideration of the noncentral nature of the distribution. Further details on calculating noncentral effect size CIs were given by Steiger (2004). However, it may not always be possible or necessary to adjust for noncentrality: Bird (2002, p. 204) observed that where the effect is not too large (e.g., $d \leq 2$) and there are sufficient degrees of freedom in the error term (more than 30), the adjustment makes little difference.

Table 10
Example Contingency Table

Seminar attendance	Adequate answer	Poor answer	Total
Attended	75	15	90
Not attended	24	30	54
Total	99	45	144

Table 11
Altered Example Contingency Table With $L = 0$

Seminar attendance	Adequate answer	Poor answer	Total
Attended	75	15	90
Not attended	45	9	54
Total	120	24	144

CIs for d can be estimated with the procedure from Grissom and Kim (2005, pp. 59–60); this estimate does not adjust for noncentrality but is useful for normally distributed data, reasonable sample sizes (at least 10 per group), and values of d that are not very large. The calculation is based on Hedges and Olkin's (1985) formula for calculating the variance (s_d^2) for the theoretical sampling distribution of d:

$$s_d^2 = \frac{n_a + n_b}{n_a n_b} + \frac{d^2}{2(n_a + n_b)},$$

where n_a and n_b are the sample sizes. The limits of the 95% CI would be

$$95\%\ CI\Delta = d \pm z_{.025} s_d.$$

Most statistics textbooks and websites provide tables of areas under the normal distribution that provide values for z at the desired cutoff. The cutoff is simply half of the difference between 1.00 and the desired CI. For a 95% CI, the cutoff is half of (1.00–.95) which is .025; table lookup provides the corresponding z value, which is 1.96. Grissom and Kim provided the following example: For $n_a = n_b = 20$ and $d = 0.7$, then $s_d^2 = 0.106$ and $s_d = 0.326$; the 95% CI would be $0.7 \pm (1.96 \times 0.326)$, giving a lower limit of 0.06 and an upper limit of 1.34. The resultant range of values for d—from almost zero to a very large effect size—is so broad that it would be difficult to draw any conclusions on the basis of the research, despite having observed a moderately large effect. Although effect sizes are independent of sample size, their presumed accuracy is increased by larger sample sizes, so the range of values in the CI becomes narrower with larger samples. If this example involved groups of 100 cases rather than 20, the bounds of the 95% CI would be .41 to .99. Replicability, as always, is an important source of confidence and even the broad ranges are useful in meta-analyses (e.g., Borenstein et al., 2009); they allow a clear pattern to emerge from multiple studies in forest plots, a

useful graphical aspect of meta-analysis (for notes about their origin, see Lewis & Clarke, 2001).

Cohen et al. (2003, p. 88) described a method for estimating CIs for R^2, provided that the sample size is greater than 60. The standard error of R^2 is calculated

$$SE_{R^2} = \sqrt{\frac{4R^2(1 - R^2)^2(n - k - 1)^2}{(n^2 - 1)(n + 3)}},$$

where n is the number of cases and k is the number of independent variables. The bounds of a 67% CI can be estimated as $R^2 \pm SE_{R^2}$; factors of 1.3, 2, or 2.6 can be applied to the standard error to provide estimates of 80%, 95%, or 99% CIs, respectively. This estimate does not adjust for noncentrality, but with larger samples, the expected error is small.

Translating Between Effect Sizes

We have described many ways of estimating effect sizes. Perhaps one of the reasons why effect sizes are underreported and infrequently discussed is that effect sizes may be reported using one statistic in one study and a different statistic in another study, making it difficult to compare the effect sizes. Many of the effect size estimates can be converted to other estimates. In Table 8, we have provided formulas for translation between d, r, and η^2.

Interpreting Effect Sizes

The object of reporting effect sizes is to better enable the reader to interpret the importance of the findings. All other things being equal, the larger an effect size, the bigger the impact the experimental variable is having and the more important the discovery of its contribution is.

In Table 7, we offer not only corresponding values for d, r, r^2, and η^2 but also two statistics—probability of superiority (*PS*) and the percentage of nonoverlap of the distributions (U_1)—that help to clarify the relationships between the distributions of the conditions being compared. The values of these statistics help the readers of reports to imagine the relationships between the two distributions from which the effect size was calculated. We suggest that one of these statistics be given along with the effect size estimate for the more important results reported in an article.

PS gives the percentage of occasions when a randomly sampled member of the distribution with the higher mean will have a higher score than a randomly sampled member of the other distribution. The values in Table 7 were abstracted from Grissom (1994). *PS* is also known as the common language effect size (McGraw & Wong, 1992). Consider, as an example, a medium size effect of $d = 0.5$ as defined by Cohen (1988). The *PS* for a d of 0.5 is 64%. That is, if you sampled items randomly, one from each distribution, the one from the condition with the higher mean would be bigger than that from the other condition for 64% of the pairs. A real-world example is given by McGraw and Wong (1992): The d for the difference in height between men and women is 2.0 for which the *PS* is 92%. That implies that if you compared randomly chosen men and women, the man would be taller than the woman for 92% of the comparisons. Finally, selecting an example from the *JEP: General* articles that we reviewed earlier, Elliot et al. (2010, Experiment 2) found that women rated men seen in pictures with a red background as more attractive than men seen against a white background, $d = 1.31$. Consulting Table 7 gives a *PS* of 82% for this value of d. That is, if pairs of pictures, one with a red and one with a white background, were selected at random, the picture with the red background would be reported as more attractive on 82% of comparisons. This use of the *PS* statistic helps to demonstrate the size of the effect in a more concrete and meaningful way than the standardized difference. This concept has been elaborated and extended by Vargha and Delaney (2000) to include all types of ordinal and interval data.

Table 7 also reports U_1, which was devised by Cohen (1988). U_1 describes the degree of nonoverlap between the two population distributions for various values of the effect sizes. For example, when $d = 0$, the populations for the two distributions are perfectly superimposed on each other, and the value of U_1 is zero; when $d = 0.5$, $U_1 = 33\%$, one third of the areas in the distributions do not overlap. $U_1 = 81\%$ for the difference between the height of men and women with $d = 2.0$ (McGraw & Wong, 1992); that is, 81% of the distributions for men and women do not overlap. For Elliot et al.'s (2010, Experiment 2) data on the attractiveness of men seen with red or white backgrounds, the U_1 percentage nonoverlap of the distributions for the value of $d = 1.31$ is 65%. As for *PS*, the U_1 statistic helps the reader to visualize the size of the effect being reported.

The substantive significance, or importance, of an effect depends in part on what is being studied. Rosnow and Rosenthal (1989), for example, illustrated how a very small effect relating to life-threatening situations, such as the reduction of heart attacks, is important in the context of saving lives on a worldwide basis (see Table 12 and Ellis, 2010). When the data are the correlation of two binary variables—such as having or not having a heart attack when in a treatment or a control condition—Rosnow and Rosenthal recommended the use of what they called the binomial effect size display to represent the relationship. The use of the binomial effect size display is illustrated in their example: Table 12 shows the frequency of heart attacks in a large study of doctors who took either aspirin or a placebo for the effect size $r = .034$. The success rate for the treatment is $.50 + r/2$ and for the control group is $.50 - r/2$. For the example in Table 12, these values are $.50 + .017$ and $.50 - .017$. The table cells are then made up to complete 100% for the columns and rows. The success rate for the treatment is calculated by subtracting the treatment effect (e.g., for aspirin) from the control effect (e.g., the placebo). For our example, that is, $51.7 - 48.3 = 3.4\%$ or $r = .034$; thus, 34 people in 1,000 would be spared heart attacks if they regularly took the appropriate dose of aspirin. It should be noted that although the simplicity of

Table 12
Binomial Effect Size Display for the Effect of Aspirin on Heart Attack Risk ($r = .034$)

Treatment	Heart attack	No heart attack	Total
Aspirin	48.3	51.7	100
Placebo	51.7	48.3	100
Total	100	100	200

Note. Values are percentages. Adapted from "Statistical Procedures and the Justification of Knowledge in Psychological Science," by R. L. Rosnow & R. Rosenthal, 1989, *American Psychologist, 44,* p. 1279. Copyright 1989 by the American Psychological Association.

calculation and clarity of presentation of the binomial effect size display is attractive, Hsu (2004) has shown that it can overestimate success rate differences unless various conditions are met.

Considerations When Reporting and Using Effect Sizes

Effect sizes estimates are important and useful descriptive statistics. Like all good descriptive statistics, they reflect the properties of the data and the conditions under which the data were collected. Just as means are valuable estimates of central tendency that can, nevertheless, be misleading if the distribution is skewed—for example, when studying income or life expectancy—so effect sizes must be considered within the context of the design and procedure, also considering the properties of the distributions. If the measures used are unreliable or if their range has been restricted, then the value of the effect size estimate will be different from, and probably smaller than, one that comes from very reliable measures or data that cover the full range. The allocation of observed variability to identified effects or to error will also influence estimates of effect size. Imagine studies of a factor that has a similar effect on people from various economic classes. One study samples only middle-class people; the error variability in this case would be smaller than the error variability in another, similar study that samples more widely. Because the error variability is smaller in the first case, the size of the effect is likely to appear larger. Yet, another study might account for variability associated with socioeconomic group by including income as a factor or covariate in the analysis, thereby reducing the error variability and increasing the apparent effect size. In general, if variables are controlled in a study and therefore do not contribute to error variability, the estimated effect size is likely to be larger than effect sizes for studies in which variables have not been controlled or have been counterbalanced across the conditions (without including the counterbalancing factor in the analysis). It is possible to correct some effect size estimates for some of these distorting factors by using statistics such as η^2_G and ω^2_G (Baguley, 2009; Grissom & Kim, 2005; Olejnik & Algina, 2003), but in all cases, interpretation and comparison of effect sizes requires careful consideration of the sources of variability.

The key point is that all estimates of effect size should be evaluated in the context of the research. It is not sensible to say of some phenomenon that its effect size is *X* without qualifying under what conditions it has been found to be *X*. Nevertheless, estimates of effect size provide both an invitation to further, meaningful interpretation and a useful metric for considering multiple, varied studies together. Complete effect size information, including the CIs of the effect size estimates, is helpful to subsequent meta-analyses, and these meta-analyses make an excellent contribution to furthering the understanding of psychological phenomena. Just as psychology researchers have become sophisticated in dealing with the complexities of inferential statistics, the regular consideration of effect sizes can lead to these statistics being demystified and becoming valuable tools.

In our surveys of the reporting of effect sizes, we have not encountered any occasion when more than one effect size was reported for any particular effect. This selectivity may result from efforts toward conciseness in reporting, or it may reflect a strategy of doing the minimum required to placate reviewers and editors. Nevertheless, we suggest that in some cases, in addition to reporting *PS* and/or U_1 to clarify the interpretation of an effect size, it is often worthwhile to report more than one measure of effect size to better interpret the results. It would, for example, be appropriate to report η^2_p to indicate the proportion of variability associated with a factor when all others are controlled, but also to report η^2_G to give an idea of the contribution that the factor makes to the overall performance when other nonmanipulated variables are allowed to vary. Both of these values would be useful in evaluating the effect. To provide another example, Cohen's *d* is useful for conceptualizing and comparing the size of a difference independently of the specific measure used; it enables comparisons between studies concerned with the same factor but using different dependent measures. However, interpretation of the results and of comparisons could be enriched by also considering *r* or r^2 as measures of the relative impact that the factor has on the outcome as is sometimes done in regression analyses where both the value of the standardized regression coefficient and the proportion of variance accounted for are discussed. The *APA Publication Manual* (APA, 2010, p. 34) specifically suggests that it will often be useful to report and discuss effect size estimates in terms of the original units, as well as the standardized approaches. The effect size expressed in original units is often clear and easy to discuss in the context of a single study, whereas the standardized units approach facilitates comparisons between studies and meta-analyses. It is also useful, when disentangling the effect of a factor with more than two levels, to provide an effect size estimate for the full effect of the factor and for each of the pairwise comparisons or other linear contrasts (see Keppel & Wickens, 2004). Similarly, analysis of simple main effects associated with an interaction should include effect size estimates both for the interaction and for the simple main effects.

Good practice with respect to effect size reporting appears to be on the increase but does not seem to have been fully adopted by most authors. Roughly half of the ANOVA reports included a measure of effect size, although few included effect size estimates for further analyses related to the ANOVA. In a few articles, authors were thorough in reporting η^2_p for the main effect in an ANOVA and reporting Cohen's *d* or η^2 for simple effects or planned or post hoc comparisons. As with all analyses, it is important to think carefully about which type of effect size is most useful for each comparison (e.g., η^2 or η^2_p). Keppel and Wickens (2004) and Rosenthal et al. (2000) provided helpful advice on using contrasts and comparisons in ANOVA designs.

We began this research with an interest in the use of effect sizes as a way of quantitatively describing effects—as a supplement to the descriptions of the data and the results of statistical tests for those effects. We found that authors have begun to include reports of effect size estimates with substantial encouragement from the APA and related professional organizations as well as from journal editors. Nevertheless, although slightly more than half of the *articles* report some effect size estimate, the majority of individual *effects* that are tested and reported are still not accompanied by descriptions of effect size. We also found that descriptions of data were often lacking. For a reader to engage with, think though, and fully consider the implications of the results of a study, descriptions of data and of the size of observed effects—both significant and nonsignificant—are needed. It is not enough to simply identify that some effects were significant and others were not. There have

been calls from some quarters to shift the emphasis away from inferential testing and toward a more descriptive and thoughtful approach to interpreting results (e.g., Cohen, 1994; Loftus, 1996). Although we are sympathetic to many of those concerns, we value an approach that includes complete reporting of statistical tests combined with descriptions of both the data and the effects. The value of a piece of research goes beyond its significant effects. The richness of the story and the argument presented by the research is essential to the development of greater understanding (e.g., Abelson, 1995), but the patterns in the data and in the effects must be reported in order for the reader to engage with the author in comprehending and evaluating the results of the research. At the moment, for most authors, considering effect sizes seems to be the last stage of their examination of their data. We believe that it should become one of the first stages. A clear grasp of the size of the effects observed is at least as important as significance testing or the calculation of CIs.

When reporting requirements change, it is usually necessary for people to learn, perhaps to teach themselves, about the new systems. It is not always easy to do so. Because we teach statistics as well as conduct research, we have been driven to explore the types of effect sizes and the usefulness of each. Our review of the reporting of effect sizes suggests that many authors have sought the minimum engagement with effect sizes that is possible while still being published. This approach is suggested by the frequent choice of effect size measures that are easily available (e.g., η_p^2) but less than optimally useful and usually not those recommended by the authors of statistical textbooks (e.g., ω^2). Statistical texts likely to be accessed by researchers are often selective in their advice about effect sizes. There are excellent discussions of the complexities of effect size available in specialist journals, but they tend to be presented in the often dense language of statistical formulas that are understandably avoided by all but the most competent or desperate researchers. We hope that this article provides a shortcut in the process of accumulating the necessary expertise to report and use effect sizes more effectively and helps people to appreciate the value of incorporating good descriptions of data and effect sizes in their reports.

We end with a minimum set of recommendations that are designed for the novice effect size user (we include ourselves in this category) and are not intended to constrain the fuller use of alternative techniques. We suggest the following:

1. Always describe the data: (a) report means or other appropriate measures of central tendency to accompany every reported analysis, and (b) report at least one associated measure of variability for each mean and the *MSE* for ANOVA analyses.

2. Also describe the effects: (a) report an effect size estimate for each reported analysis, (b) for the most important effects, report complete effect size information, including the CIs of the effect size estimates for possible use in subsequent meta-analyses, (c) for the difference between two sets of data, as a default, use Cohen's *d* (or Hedges's *g*) as the effect size estimate and, for small sample sizes, also report $d_{unbiased}$, and (d) for factorial analyses, with due thought and consideration, select and report η^2, η_G^2, and/or η_p^2 as appropriate for the interpretation provided in the report, and where effects or *N*s are small, indicate the possible inflation of η^2 by also reporting ω^2, ω_G^2, and/or ω_p^2.

3. For complex analyses, such as factorial ANOVA or multiple regression, report all effects. Report the results for each effect, including *F, df,* and *MSE*, so that the reader can calculate effect sizes other than those reported.

4. Take steps to aid the reader to understand and interpret the size of the more important effects. Use statistics such as the *PS* and Cohen's U_1 or Goodman–Kruskal's lambda to help the reader conceptualize the size of the effect.

5. Always discuss the practical, clinical, or theoretical implications of the more important of the effect sizes obtained.

References

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

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

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Aron, A., Aron, E. N., & Coups, E. (2009). *Statistics for psychology* (5th ed.). Upper Saddle River, NJ: Pearson.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100,* 603–617. doi:10.1348/000712608X377117

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37,* 379–384. doi:10.3758/BF03192707

Becker, L. A. (2000). Effect size calculators. Retrieved from http://www.uccs.edu/~faculty/lbecker/

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement, 62,* 197–226. doi:10.1177/0013164402062002001

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* Chichester, UK: Wiley. doi:10.1002/9780470743386

Campbell, J. P. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67,* 691–700. doi:10.1037/h0077946

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153. doi:10.1037/h0045186

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003. doi:10.1037/0003-066X.49.12.997

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.

Coolican, H. (2009). *Research methods and statistics in psychology.* London, United Kingdom: Hodder.

Cramér, H. (1946). *Mathematical methods of statistics.* Princeton, NJ: Princeton University Press.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York, NY: Routledge.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement, 61,* 532–574.

Elliot, A. J., Niesta Kayser, D., Greitemeyer, T., Lichtenfeld, S., Gramzow, R. H., Maier, M., & Liu, H. (2010). Red, rank, and romance in women viewing men. *Journal of Experimental Psychology: General, 139,* 399–417. doi:10.1037/a0019689

Ellis, P. D. (2009). Effect size calculators. Retrieved from http://myweb.polyu.edu.hk/~mspaul/calculator/calculator.html

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results.* Cambridge, United Kingdom: Cambridge University Press.

Ezekiel, M. (1930). *Methods of correlational analysis.* New York, NY: Wiley.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79,* 314–316. doi:10.1037/0021-9010.79.2.314

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach.* New York, NY: Psychology Press.

Grissom, R. J., & Kim, J. J. (2011). *Effect sizes for research: A broad practical approach* (2nd ed.). New York, NY: Psychology Press.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York, NY: Holt, Reinhart, & Winston.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92,* 490–499. doi:10.1037/0033-2909.92.2.490

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

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician, 55,* 19–24. doi:10.1198/000313001300339897

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Hsu, L. M. (2004). Biases of success rate differences shown in binomial effect size displays. *Psychological Methods, 9,* 183–197. doi:10.1037/1082-989X.9.2.183

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62,* 227–240. doi:10.1177/0013164402062002002

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in planning estimation via narrow confidence intervals. *Psychological Methods, 11,* 363–385. doi:10.1037/1082-989X.11.4.363

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson.

Levant, R. F. (1992). Editorial. *Journal of Family Psychology, 6,* 3–9. doi:10.1037/0893-3200.6.1.5

Lewis, S., & Clarke, M. (2001). Forest plots: Trying to see the wood and the trees. *British Medical Journal, 322,* 1479–1480. doi:10.1136/bmj.322.7300.1479

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5,* 161–171. doi:10.1111/1467-8721.ep11512376

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59,* 537–563. doi:10.1146/annurev.psych.59.103006.093735

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11,* 386–401. doi:10.1037/1082-989X.11.4.386

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361–365. doi:10.1037/0033-2909.111.2.361

Morris, P. E., & Fritz, C. O. (2011). *The reporting of effect sizes in cognitive publications.* Manuscript submitted for publication.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82,* 3–5. doi:10.1037/h0092448

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

Richardson, J. T. E. (1996). Measures of effect size. *Behavior Research Methods, Instruments & Computers, 28,* 12–22. doi:10.3758/BF03203631

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

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284. doi:10.1037/0003-066X.44.10.1276

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

Smithson, M. (2003). *Confidence intervals.* Thousand Oaks, CA: Sage.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9,* 164–182. doi:10.1037/1082-989X.9.2.164

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson.

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51,* 473–481. doi:10.1037/0022-0167.51.4.473

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

Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics, 2,* 440–457. doi:10.1214/aoms/1177732951

Wilkinson, L., & the APA 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

Yates, F. (1951). The influence of statistical methods for research workers on the development of the science of statistics. *Journal of the American Statistical Association, 46,* 19–34. doi:10.2307/2280090

(*Appendix follows*)

Appendix

A Brief Introduction to Power Analysis

When planning research, it is sensible to ensure that the research has the power to detect the effect(s) under consideration (e.g., Ellis, 2010; Keppel & Wickens, 2004). The anticipated size of the effect may be estimated from previous research, or an effect size may be chosen that is the smallest effect that would be meaningful in some practical sense. It is also good practice to define the degree of power required, that is, to set an acceptable probability of Type II errors. Although the limit for Type I errors is usually set to .05, the limit for Type II errors is often set to be somewhat higher; if the limit for Type II errors is set to .20, then power would need to be .80, a commonly recommended value (Ellis, 2010). Having identified an anticipated effect size and the power requirements, it is easy to determine the number of participants required. Table A1 is adapted from Cohen (1988); the intersection of the selected effect size and power level shows the number of participants required in each group for both two-tailed and one-tailed *t* tests at a significance threshold of $\alpha = .05$. Three levels of effect size are included in this brief summary: small ($d = .2$), medium ($d = .5$), and large ($d = .8$); Cohen's tables provide data for a fuller range of effect sizes. If the planned research will require too many participants to be practicable, then it may be worthwhile to consider ways of reducing error variability, thereby increasing the anticipated effect size.

When an experiment has led to a nonsignificant result, it is inappropriate to post hoc calculate the power of the study (Ellis, 2010; Hoenig & Heisey, 2001). However, the experiment can provide an estimate of the population effect size, although this may not be very accurate if the sample size was small. Using the estimate of the population effect size, future research can be planned with respect to the sample size required to achieve a reasonable level of power.

Table A1
Number of Participants per Group Required for t Tests to Achieve Selected Levels of Power, Based on the Anticipated Size of the Effect

	Effect size for one-tailed test			Effect size for two-tailed test		
Power	Small ($d = .2$)	Medium ($d = .5$)	Large ($d = .8$)	Small ($d = .2$)	Medium ($d = .5$)	Large ($d = .8$)
.25	48	8	4	84	14	6
.50	136	22	9	193	32	13
.60	181	30	12	246	40	16
.67	216	35	14	287	47	19
.70	236	38	15	310	50	20
.75	270	44	18	348	57	23
.80	310	50	20	393	64	26
.85	360	58	23	450	73	29
.90	429	69	27	526	85	34
.95	542	87	35	651	105	42

Note. Where power is .8, there is a 20% chance of failing to detect an effect. Adapted from *Statistical Power Analysis for the Behavioral Sciences* (2nd ed., pp. 54–55), by J. Cohen, 1988, Hillsdale, NJ: Erlbaum. Copyright 1988 by Taylor & Francis.

Table A2
Power Present Based on the Number of Groups, Number of Participants per Group, and Meaningful or Expected Effect Size

	Effect size		
Participants per group	Small ($d = .2$; $\eta^2 = .01$)	Medium ($d = .5$; $\eta^2 = .06$)	Large ($d = .8$; $\eta^2 = .14$)
Two groups			
10	.07	.18	.40
15	.08	.26	.57
25	.10	.42	.80
40	.14	.61	.95
80	.24	.89	1.00
Three groups			
10	.07	.20	.45
15	.08	.29	.64
25	.10	.47	.87
40	.15	.68	.98
80	.27	.94	1.00
Four groups			
10	.07	.21	.51
15	.08	.32	.71
25	.11	.53	.93
40	.16	.76	.99
80	.29	.97	1.00
Five groups			
10	.07	.23	.56
15	.09	.36	.78
25	.12	.58	.96
40	.17	.81	1.00
80	.32	.99	1.00

Note. These figures apply to analysis of variance and to two-tailed t tests. Where power is .3, there is a 70% chance of failing to detect an effect. Adapted from *Statistical Power Analysis for the Behavioral Sciences* (2nd ed., pp. 311–318), by J. Cohen, 1988, Hillsdale, NJ: Erlbaum. Copyright 1988 by Taylor & Francis. Cohen's tables report another effect size estimate, *f*, which is rarely reported and is not addressed in this article. The relationship between *f* and η^2 is $f = \sqrt{\frac{\eta^2}{1-\eta^2}}$ and $\eta^2 = \frac{f^2}{1-f^2}$.

Table A2 is also adapted from Cohen (1988); it lists power levels for small, medium, and large effect sizes given some number of groups and participants per group. These values apply to ANOVAs and two-tailed *t* tests.

Power may not be the sole consideration when estimating the number of participants required. Sample means and variability provide estimates of the population parameters; the accuracy or precision of those estimates is a function of the sample size. It may be as useful or more useful in some cases to estimate the sample size required for a desired degree of accuracy in parameter estimation based on defining the maximum acceptable confidence interval width. Maxwell, Kelley, and Rausch (2008) provided an excellent discussion of power and accuracy in parameter estimation; practical guidance is available there and in other articles (e.g., Kelley & Rausch, 2006) and texts (Cumming, 2012).

Received April 1, 2011
Revision received May 15, 2011
Accepted May 15, 2011 ■

Correction to Fritz et al. (2011)

The article "Effect Size Estimates: Current Use, Calculations, and Interpretation," by Catherine O. Fritz, Peter E. Morris, and Jennifer J. Richler (*Journal of Experimental Psychology: General*, Advance online publication. August 8, 2011. doi:10.1037/a0024338) contained a production-related error. The sixth equation under "Effect Sizes Specific to Comparing Two Conditions" should have had a plus sign rather than a minus sign in the denominator. All versions of this article have been corrected.

DOI: 10.1037/a0026092