British Journal of Psychology (2009), 100, 603–617
© 2009 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Standardized or simple effect size: What should be reported?

Thom Baguley*
Division of Psychology, Nottingham Trent University, Nottingham, UK

It is regarded as best practice for psychologists to report effect size when disseminating quantitative research findings. Reporting of effect size in the psychological literature is patchy – though this may be changing – and when reported it is far from clear that appropriate effect size statistics are employed. This paper considers the practice of reporting point estimates of standardized effect size and explores factors such as reliability, range restriction and differences in design that distort standardized effect size unless suitable corrections are employed. For most purposes simple (unstandardized) effect size is more robust and versatile than standardized effect size. Guidelines for deciding what effect size metric to use and how to report it are outlined. Foremost among these are: (i) a preference for simple effect size over standardized effect size, and (ii) the use of confidence intervals to indicate a plausible range of values the effect might take. Deciding on the appropriate effect size statistic to report always requires careful thought and should be influenced by the goals of the researcher, the context of the research and the potential needs of readers.

There is now near universal agreement in the psychological literature that reports of statistical procedures such as null hypothesis significance tests should be accompanied by an appropriate measure of the magnitude of the effect (e.g. Abelson, 1995; Wilkinson & APA Task Force on Statistical Inference, 1999). Reporting effect size aims to facilitate: (i) understanding of the importance of an effect – in particular its practical importance (see Kirk, 1996), (ii) comparison of effect sizes within or between studies, and (iii) secondary analysis (e.g. power calculations or meta-analysis).

The practice of reporting effect size is complicated, however, by the large number of different measures of effect size from which to select. There is a growing literature on what measure ought to be selected (e.g. Kirk, 1996; Olejnik & Algina, 2000, 2003), but it would be unrealistic to expect many researchers to keep up with the full range of available effect size metrics. The aim of this paper is to consider how best to report effect size, with particular focus on the choice between *standardized* and *simple* effect size.

*Correspondence should be addressed to Dr Thom Baguley, Division of Psychology, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, UK (e-mail: Thomas.Baguley@ntu.ac.uk).

DOI:10.1348/000712608X377117

Standardized measures of effect size

A standardized measure of effect is one which has been scaled in terms of the variability of the sample or population from which the measure was taken. In constrast, simple effect size (Frick, 1999) is unstandardized and expressed in the original units of analysis. Rosenthal (1994) classifies standardized effect sizes into one of two main families: the *r* family and the *d* family. An important distinction between *r* and *d* is that in a two-group independent design when both are applicable, *d* but not *r* is not sensitive to the base rates (N_1 and N_2) of the groups (McGrath & Meyer, 2006). The *r* family includes Pearson's *r* and variants such as r^2 or Fisher's *z* transformation. The *d* family includes standardized mean differences such as Cohen's *d* and Hedge's *g*. Within each family measures may be descriptive (e.g. *d* or η^2 that reflect variance explained in a sample) or inferential (e.g. *g* or ω^2 that estimate population parameters). In order to properly appreciate the distinction between standardized and simple measures of effect it is important to consider how measures such as *r* or *d* are computed.

The starting point for Cohen's *d* is a simple effect size metric: the simple difference between the means being compared: $M_1 - M_2$ (e.g. the experimental group mean minus the control group mean). Standardization is achieved by dividing the difference $M_1 - M_2$ by a standard deviation (*SD*) – usually the pooled *SD* (σ_{pooled}) of the scores that contribute to the mean. Although other members of the *d* family use variants of $M_1 - M_2$ as the numerator (e.g. Morris & DeShon, 2002) or alternatives to σ_{pooled} as the denominator (see Rosenthal, 1994), what they share is that they scale a simple difference between means in *SD* units. In other words $d = 1$ represents a 1 *SD* difference in the means.

An *r* value can be thought of in much the same way. Consider a simple linear regression between *X* and *Y*. The original values of *X* and *Y* may be standardized by replacing them by *z* scores (i.e. by subtracting the mean of *X* or *Y* from each score and dividing the result by the *SD* of *X* or *Y*). Linear regression of *X* and *Y* expressed as *z* scores produces a standardized coefficient, β, for the slope of the regression line. In the case of bivariate linear regression, β is identical to *r*. Just as the unstandardized slope of a regression line can be interpreted as the number of units of increase in *Y* associated with an increase of 1 unit in *X*, *r* (or β) is the number of *SD*s we expect *Y* to increase for each *SD* increase in *X*.

These examples should make it clear that both *r* and *d* take an effect in the original units of analysis and transform them by replacing those original units with the *SD*. Other standardized measures operate in a similar way. Thus measures of 'variance explained' such as r^2 standardize using the variance (SD^2). The rationale for using such measures is intuitively appealing (but potentially dangerous): we can replace the original units with common units that supposedly facilitate comparison. Thus, the decision to report standardized effect size in place of simple effect size is, in essence, a decision about whether to switch from the original units to the *SD*.

Difficulties arising from standardization

The principal aim of standardization is to equate effects measured on different scales. It is not clear that standardization is successful in this aim. For example, two studies reporting *d* may well compute the statistic with different choices of *SD* unit (Morris & DeShon, 2002). A highly desirable property in an effect size measure would be that it remain stable between different versions of the same measurement instrument, between individuals scoring high or low on one of the variables, or between different

study designs. Standardized effect size is particularly vulnerable to changes in any of these factors, because all three influence sample variance.

Reliability

In a statistical model we can consider a parameter estimate (such as a mean) as a sum of its true value plus error. The error term in the model can in turn be broken down into other components (such as individual differences between the people or units being measured). One component in the error term is the measurement error associated with a sample of scores – though this itself can be partitioned into different sources of error (Schmidt & Hunter, 1999). Amongst other things, it will vary with the precision of the scores obtained from a measurement instrument (e.g. measuring height with a ruler is less precise than with a tape measure). Studies using two versions of the same instrument, such as the short and long form of a psychometric scale, usually differ in reliability (even if all other sources of error are held constant). This will produce spurious differences in standardized effect size statistics such as r or d.

The influence of reliability on effect size depends on the nature of the statistical model that is employed. In a simple model with a single predictor and a single outcome unreliability of X or Y will always reduce standardized effect size. Unreliability also always reduces standardized effect sizes in ANOVA models where all the factors are orthogonal. This is because unreliability inflates the estimate of variability in the population of interest and exaggerates the size of the *SD* or variance used to scale the effect (Ree & Caretta, 2006). It should also be noted that in some non-orthogonal designs it is reasonable to assume that all X variables are measured with perfect or near-perfect reliability (e.g. for predictors such as gender). If so, unreliability of Y will likewise depress standardized effect size. However, in complex designs with correlated predictors that differ in reliability both standardized *and* simple effect sizes estimates may be distorted (Ree & Caretta, 2006).

Corrections for reliability

The effect of reliability on correlation is shown by the attenuation formula in classical measurement theory (e.g. Ghiselli, 1964):

$$r_{xy} = r_{x_t y_t} \sqrt{\left(r_{xx} r_{yy}\right)} \tag{1}$$

This shows that the observed correlation between X and Y, r_{xy}, is a function of the 'true' correlation in the population sampled, $r_{x_t y_t}$, and the reliability with which X and Y are measured (r_{xx} and r_{yy}). In this simple case, we can simply rearrange the attenuation formula to disattenuate the correlation for the effects of reliability:

$$r_{x_t y_t} = \frac{r_{xy}}{\sqrt{\left(r_{xx} r_{yy}\right)}} \tag{2}$$

Bobko, Roth and Bobko (2001) report an equivalent, but less well known, correction for use with d:

$$d_{\text{corrected}} = \frac{d_{\text{observed}}}{\sqrt{r_{yy}}} \tag{3}$$

This correction corrects only for the reliability of *Y* (and thus assumes that the dichotomous *X* variable is measured without error). Few researchers (outside specialist applications such as meta-analysis) correct for measurement error (Bobko *et al.*, 2001). Many researchers are unaware of the desirability of such corrections, or collect data for which the reliability of some measures is unknown (or hard to obtain). Even when appropriate corrections are employed researchers frequently use reliability estimates that do not take into account all potential sources of measurement error and thus tend to 'undercorrect' (Schmidt & Hunter, 1999). It is also possible to overcorrect by applying the wrong reliability estimate. If some sources of error contained in the reliability estimate are an intrinsic aspect of the effect of interest (e.g. if the effect is changing over time) such errors are particularly difficult to avoid (DeShon, 2003).

Range restriction

Standardized effect size is also influenced by the way people (or other units of analysis) are sampled. If the sample is restricted to a subset of the population of interest this will influence the variance. Sampling from a truncated distribution (missing one or both tails) will reduce the *SD*. Sampling only the tails will increase the *SD*.[1] So selecting participants who score above some threshold on a criterion (e.g. extraversion) will lower the *SD*. If what is measured correlates with this criterion the covariance between *X* and *Y* will also decrease. This tends to reduce the sample correlation relative to the 'true' value in the unrestricted population. To illustrate this, consider the relationship between *r* and the unstandardized slope, *b*, in a regression:

$$r_{xy} = b_{xy}\frac{SD_x}{SD_y} \tag{4}$$

It follows (except when $b_{xy} = 0$) that if something reduces SD_x relative to SD_y then it will decrease *r*. If *X* and *Y* are correlated then range restriction on *X* tends also to decrease SD_y but the expected decrease in SD_x always exceeds that of SD_y unless $r = 1$.

A corollary of range restriction is that sampling the extremes or tails of the criterion variable (i.e. avoiding the middle) will tend to increase the observed correlation with *Y*. This strategy of sampling extreme values is a common way to increase the power of a study, but it is rarely appreciated that it also inflates standardized effect size (Preacher, Rucker, MacCallam, & Nicewander, 2005). Figure 1 (a) and (b) show how sampling the extremes of *X* has negligible impact on an unstandardized slope, but increases the slope of the standardized slope (d) relative to that of full data set (c).

Range restriction is a common consequence of any selective sampling strategy. Such selection often occurs inadvertently (e.g. if sampling a subpopulation with a mean higher or lower than the overall population). Range restriction can also arise after data collection through practices such as trimming or 'outlier' deletion (Wright, 2006).

[1] *If this relationship does not seem obvious, recall that the* SD *is calculated using the squared distances from the mean. For a linear relationship we would expect the mean to be in a similar location whether the tails or the middle of the distribution are sampled. Sampling from the tails inevitably increases the distance from the mean and hence the* SD. *Excluding the tails decreases distance from the mean and reduces the* SD.

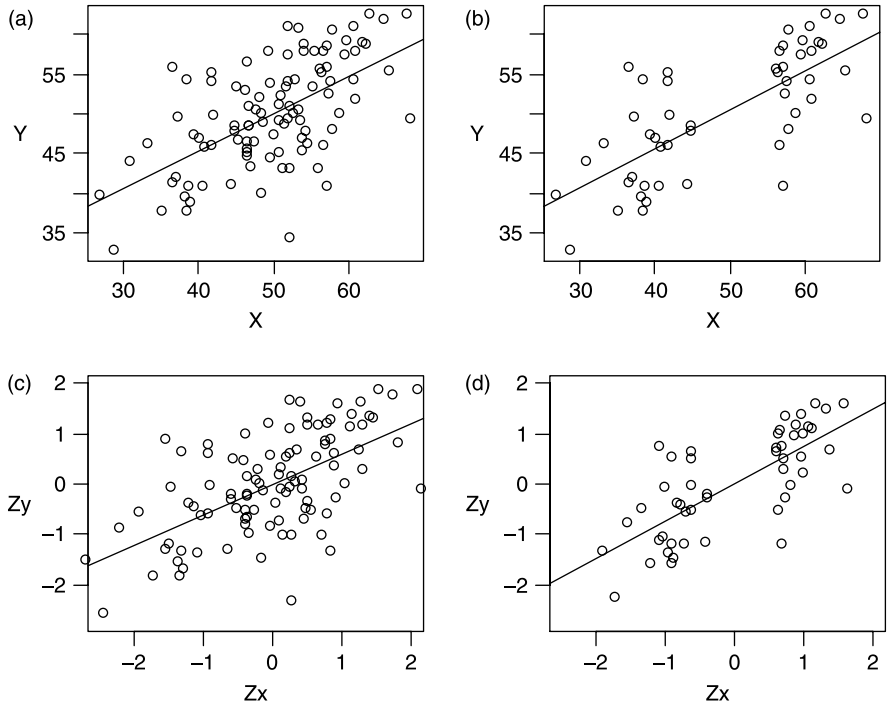

Figure 1. The corollary of range restriction: sampling the extremes of *X* has negligible effect on the unstandardized slope, but increases the standardized slope for the regression of *Y* on *X*. (a) The unstandardized slope between two normal, random variables: *X* and *Y*; $Y = 26.34 + 0.4743X$. (b) The unstandardized slope, selecting only the upper and lower quartiles of *X*; $Y = 26.10 + 0.4894X$. (c) The standardized slope of *X* and *Y* ($r_{99} = .605$). (d) The standardized slope of *X* and *Y* selecting only the upper and lower quartiles of *X* ($r_{49} = .735$).

Corrections for range restriction

Correcting for range restriction is even less widely practised than correcting for reliability. The correction can be illustrated with the case of simple regression or correlation in which direct range restriction occurs on *X*, but *Y* is unrestricted (e.g. Ghiselli, 1964):

$$r_{x_t y_t} = \frac{k r_{xy}}{\sqrt{k^2 r_{xy}^2 - r_{xy}^2 + 1}} \tag{5}$$

Here $r_{x_t y_t}$ is the 'true', unrestricted correlation in the population, r_{xy} is the observed sample correlation and k is the ratio of the unrestricted to restricted variance. Similar corrections can be applied for *d* (Bobko *et al.*, 2001). It is also possible to combine corrections for reliability and range restriction, although the corrections are more complex. Furthermore, although most cases of range restriction are indirect, many researchers inappropriately apply direct range restriction equations (Hunter *et al.*, 2006).

Study design

Common standardized measures of effect size are typically not stable between studies with different designs. One illustration of this is the case of independent and repeated

measures designs. Consider a data set of reading times for four-sentence spatial descriptions adapted from Baguley and Payne (2000). The data consist of reading times per syllable (in seconds) averaged over a number of descriptions (summarized in Table 1).

Table 1. Mean and standard deviation reading times per syllable by sentence number (adapted from Baguley & Payne, 2000)

	N	*M*	*SD*
Sentence 1	71	1.200	1.000
Sentence 2	71	1.058	1.201
Sentence 3	71	0.914	0.422
Sentence 4	71	0.761	0.397

The original data were repeated measures and produce a statistically significant main effect, $F(3,210) = 5.34, p < .05, \eta_p^2 = .071$. This effect size measure, partial eta-squared (η_p^2) can be readily calculated from the ANOVA table:

$$\eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{error}}, \quad \text{or} \quad \frac{df_{\text{effect}} \times F}{(df_{\text{effect}} \times F + df_{\text{error}})} \tag{6}$$

Suppose that the same numerical values had been obtained from an independent measures design. In this case the main effect, $F(3,280) = 3.64, p < .05, \eta_p^2 = .037$, is still statistically significant, but η_p^2 is considerably lower. This is problematic because it is purely an artefact of the method by which η_p^2 is calculated: it calculates the proportion of variance explained for an effect excluding all other effects. Here the repeated measures design treats individual differences as a separate effect and therefore excludes them from the calculation. This type of problem can occur whenever any partial effect size statistic is used (not just η_p^2). Nor does it occur merely because the two designs aim to test different hypotheses. The advantage of using repeated measures designs is that, in principle, it estimates the same population parameters as the independent measures design with greater statistical power.

A further important influence of study design on standardized effect size arises from the distinction between *manipulated* and *stratified* factors (Gillett, 2003; Olejnik & Algina, 2003). A manipulated factor is an 'experimenter-devised treatment variable' such as the retention interval or length of a word list in a memory experiment. A *stratified* factor (sometimes called a *measured* or *individual difference* factor) is a partitioning of a sample into homogeneous subsets (e.g. by gender). If the variances of the subpopulations being sampled differ then an identical mean difference will (trivially) produce a larger standardized effect size for the more homogeneous group. Buchner & Mayr (2004) argue that precisely this problem has contributed to an apparent young-old difference in auditory negative priming. Unstandardized reaction times tend to show larger negative priming effects for older than younger adults (e.g. 86 ms versus 53 ms), yet because older adults have more variable responses the standardized effect is smaller for older participants than younger ones (e.g. $d = 0.58$ versus $d = 0.83$).[2] If Buchner

[2] *Buchner and Mayr (2004, Experiment 1).*

and Mayr are correct then misapplication of standardized effect size has contributed to at least one influential, but erroneous, finding in experimental psychology.

There are also more subtle consequences of the type of factor. Gillett (2003) compared one-factor (factor A manipulated alone) and two-factor (adding stratified factor B) analyses of a data set. If the subpopulations (e.g. males and females) used to stratify B have different means then *d* or *r* will be smaller in the one-factor design than the two-factor design. This happens because of the reduction in model error with the inclusion of factor B (analogous to how individual differences influence η_p^2 in a repeated measures design). For designs where only manipulated factors are employed and where participants are randomly assigned to a treatment level, things are more straight-forward. Under the assumption of *unit-treatment additivity* (i.e. that the only effect of a treatment is to add or subtract a constant to the mean) standardized effect size will remain stable. If unit-treatment additivity is violated the estimates may be distorted by the particular levels of a manipulated factor selected by the researcher. For instance, a study looking at the effect of caffeine on heart rate might expect to find that caffeine increased heart rate, but one might also expect large doses of caffeine to make heart rate more variable. If so, the effects of high doses of caffeine would be underestimated using *r* or *d*.

A particularly thorny issue concerns the role of fixed and random effects in the calculation of effect size. A fixed effect is one that is considered to sample the population of interest exhaustively, whereas a random effect is one for which a finite sample is taken from the population of interest (which is presumed to be infinite). Many statistical analyses familiar to psychologists assume that there is a single locus of error in the sample: random variation between the units of analysis (usually people). In some analyses there are additional loci of random variation that ought to be modelled (e.g. within people in a repeated measures design). Aside from repeated measures designs, the context in which most psychologists encounter this issue (if they encounter it at all) is in terms of the *language-as-a-fixed-effect fallacy* (Clark, 1973), but it can also arise in many other contexts. Clark noted that whilst psychologists routinely treat participants as a random factor in statistical designs they routinely treat samples of words as fixed. Clark argued that this is inappropriate if researchers want their findings to generalize beyond the words they sampled. A common, albeit flawed, solution (Raaijmakers, Schrijnemakers, & Gremmen, 1999) is to report separate analyses of the effects *by participants* (treating only participants as a random factor) and *by items* (treating items, but not participants, as a random factor). Each such analysis ignores a major source of sampling variability (participants or items) and arguably inflates standardized effect size. In addition, it is inappropriate to compare standardized effect sizes computed from *by items* and *by participants* analyses because they are computed using different denominators and are thus on different scales.

Adjusting for differences in study design

The main difficulty in dealing with differences in design is that the precise nature of the adjustment required to equate two effect size statistics depends both on the statistic that is used and on the comparison one wishes to make. A relatively simple case is that of a difference in means. If one calculates *d* from a paired *t* test the observed *d* will typically be much higher than the value for the equivalent independent design. For this reason Dunlap, Cortina, Vaslow, and Burke (1996) propose that the original sample *SD*s should be used to calculate *d*. However, this will not always be appropriate: an alternative conception of *d* using a change score rather than a difference score will be preferable in some situations (Morris & DeShon, 2002).

The appropriate procedure for other contexts is also difficult. We can sometimes avoid problems associated with use of η_p^2 by reporting eta-squared (η^2):

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \tag{7}$$

In the simple case of comparing one-factor repeated with independent measures ANOVA, η^2 works reasonably well. As one-factor independent measures ANOVA only partitions variance into two sources: SS_{effect} and SS_{error} (and because these must sum to SS_{total}) η_p^2 and η^2 are equivalent. Complications arise in factorial ANOVA. Consider the comparison of effects of factor A in a two-factor design (factors A and B) with a one-factor design (factor A alone). In this case η_p^2 will be preferable to η^2 provided both A and B are manipulated factors, because η_p^2 strips out the extraneous variation introduced by manipulating B. The appropriate correction to standardized effect size thus depends on type of design, the nature of the factors (manipulated or stratified) and the comparisons one wishes to make. Olejnik and Algina (2000; 2003) describe how to calculate generalized versions of statistics such as *d*, η^2 or ω^2 that are stable across a range of different designs.

The nature of the comparisons one wishes to make is particularly important in considering the appropriate adjustment for standardized effect sizes computed from studies where items may be considered a random factor. There are some contexts (e.g. theory testing) where it might be sufficient to show *by participants* or *by items* effects. In contrast, an effect size estimate that ignores substantial sources of variability in the populations of interest would be very misleading for assessing practical importance. An effect might account for substantial variation between participants (e.g.$\eta^2 = .60$) but a negligible proportion between items (e.g. $\eta^2 = .04$). A *by participants* analysis would therefore give a misleading estimate of the practical importance of the effect for items other than those sampled (for which the estimate of η^2 could not exceed .04). In this case *minF'* (Clark, 1973) might be used to derive a conservative effect size estimate, for example via Equation (6), though there appears to be no specific support for this in the literature.

The robust beauty of simple effect size

A straight-forward alternative to standardized effect size is to rely on simple effect size (e.g. a raw difference in means or an unstandardized slope coefficient). This approach has three principal advantages over standardized effect size.

The first advantage is that the scale is independent of the variance. This means that simple effect size avoids all problems that arise solely from standardization. Simple effect size is therefore far more robust than standardized effect size. Although problems with standardization are well documented (though often ignored) in relation to regression models (e.g. Tukey, 1969; Kim & Ferree, 1981; Greenland, Schliesselman, & Criqui, 1986) they are not widely known in relation to other statistical procedures. Even if standardized effect sizes are adjusted for reliability, range restriction and differences in study design those adjustments can *at best* put standardized and simple effect size on equivalent terms.

The second advantage is that, because simple effect size is scaled in terms of the original units of analysis, it will nearly always be more meaningful than standardized effect size. As Tukey (1969, p. 89) put it: 'being so disinterested in our variables that we do not care about their units can hardly be desirable.' Baguley (2004) emphasized this

point in the context of applied research – where it is likely that many consumers of research will be familiar with the interpretation of common units of measurement in their field, but less familiar with how to interpret standardized effect size metrics. (Worse still psychologists may be familiar with standardized effect such metrics such as R^2, but consistently misinterpret them because they do not understand how they are calculated and what factors influence them.)

A similar point can be made for theoretical research. Units of measurement are rarely completely arbitrary and their advantages, disadvantages and appropriate interpretation are typically part of the 'lore' of a discipline (Abelson, 1995). In general, the more that is known about the context of a study and the more familiar researchers and research consumers are with the units of measurement being used, the more useful simple effect size is as an indicator of the importance of a finding.

The argument here is not that simple effect size has a one-to-one mapping with the underlying psychological construct or constructs of interest. Rather, simple effect size retains more information about the context of the data than standardized effect size. Even if the measure is not particularly meaningful (e.g. an arbitrary rating scale with unknown psychometric properties) standardization obscures the deficiencies of the measure (Tukey, 1969) as well as limiting the application of what we do understand about the measure. Most researchers know that a two point difference on a five point rating scale is substantial (e.g. indicating a shift from an extreme position to a neutral one), whereas $d = 0.25$ might indicate a large shift in a noisy sample or a small shift in a very homogenous one.

The third advantage is a practical one. Simple effect size is easier to compute than standardized effect size. Less computation means less opportunity for computational or other errors to occur (e.g. correcting using the wrong reliability coefficient). Metrics such as η_p^2 are easy to obtain from widely used computer packages, but no computer package can automatically incorporate reliability corrections, range restriction corrections or deal with issues relating to study design. Such decisions are sensitive to specific context. For many situations, calculating an appropriate correction or adjustment is an unnecessary step that either replicates the information in simple effect size or risks introducing errors.

In addition to these principal advantages it is worth noting the relationship between simple effect size and what Abelson (1995) has termed *causal efficacy* – the ratio of the size of an effect to the size of its cause. An effect is potentially more interesting or important if it produces a relatively large effect from a relatively small cause. Simple effect size, in the form of the unstandardized slope of a regression line, is itself a measure of causal efficacy. With a little care it is very easy to recover causal information using simple effect size. For example, regression can often replace ANOVA if the levels of a factor are sampled from a continuous distribution (e.g. different delays in a memory task). Not only is the analysis likely to be more powerful, but the slope (b) will provide an easy-to-interpret estimate of the effect (e.g. the rate of forgetting). This approach is particularly useful because it strips out both the effects of range restriction illustrated in Equation (4) and the dose-effect relationship of X on Y.

Simple, unstandardized effect size eliminates many, but not all, of the problems associated with selecting and calculating an appropriate and accurate standardized effect size metric. If a statistical model is mis-specified in some way (e.g. if a confounding factor is not included, or the dose-effect relationship is not linear) *any* measure of effect size derived from that statistical model will be inaccurate. It is also possible that the original units of measurement may not be ideal (e.g. more appropriate units may involve

a transformation). A memory researcher might consider a measure of signal detection such as d' or a simpler measure such as the proportion of hits minus false alarms. In these cases an important consideration is the theoretical model being considered (as different measures may imply different models). The point made here is *not* that the original measures are always the best choice, but that simple effect size is much more robust than uncorrected standardized effect size and typically much easier to interpret than either corrected or uncorrected standardized effect size.

Reporting effect sizes for categorical data

Thus far discussion has focussed on effect sizes for continuous measures. Standardized effect size is rarely advocated for categorical outcomes – in part because some problems associated with standardized effect size are aggravated when an outcome is not continuous. A popular standardized effect size in this case is ϕ (phi): equivalent to Pearson's r between the variables in a 2×2 contingency table. For artificial categories ϕ is particularly misleading: a continuous measure behaves as if measured with extremely low reliability when dichotomized. Even if restricted to genuine categories, ϕ has undesirable properties: two tables with the same percentage outcomes but different marginal totals may produce quite different values of ϕ (Fleiss, 1994).

A full discussion of effect sizes for categorical data is beyond the scope of this article, but it seems likely that many psychologists would benefit from using *odds ratios* when reporting categorical effects (not least because they readily generalize to techniques such as logistic regression). Odds ratios are base-rate insensitive measures of effect size (McGrath & Meyer, 2006). Base-rate sensitive measures such as *risk ratios* or *number needed to treat* may be more appropriate for applications such as communicating findings in clinical settings (e.g. Walter, 2000).

Should standardized effect sizes ever be used?

There are, however, two broad situations where standardized effect size may be preferable to simple effect size: i) when the primary goal of the research is to estimate standardized effect size, and ii) when comparing conceptually similar effects using different units of measurement.

Estimating a standardized effect size is rarely the primary goal of research. In applied research users nearly always want to relate the observed effect size to the original context using the original units of measurement (Baguley, 2004). For standardized measures in the d family of effect size metrics it is difficult to imagine situations in which the primary goals of the researcher could not be met using a simple difference in means. The precise value d takes is somewhat arbitrary (with the exception of $d = 0$) and substantive questions about the magnitude of effect can be readily answered using the simple difference. For the r family there are several situations involving continuous outcomes where simple effect size may be inferior to standardized effect size. r may take the non-arbitrary values of -1, 0 and 1, while r^2 can take the non-arbitrary values of 0 and 1. If the goal of a researcher is to address a substantive hypothesis corresponding to one of these non-arbitrary values then it may be useful to focus on standardized effect size. This case is equivocal when the hypothesis is one of no effect (e.g. $r = 0$). In these cases a simple regression coefficient would provide similar information (and the original units may be more revealing in terms of practical importance). More striking is the situation when a substantive hypothesis is that $r = 1$ or -1 (or $r^2 = 1$).

It is important here to clarify what is meant by a substantive hypothesis. A substantive hypothesis is derived from theory and is somewhat plausible (e.g. it is reasonable to believe that $r^2 = 1$ or very close to 1). This definition rules out the traditional null hypothesis (H_0) in statistical testing. H_0 is usually somewhat *implausible* (and only rarely of theoretical interest). Prime examples of substantive hypotheses of this type occur in the psychometric literature where the hypothesis that a scale is highly reliable or valid is of genuine interest. For reliability, the proposition that $r_{yy'} = 1$ is of interest (as this is desired level of reliability for any measure) and it is quite possible to obtain reliabilities very close to 1. The standardized slope of the regression does lose information (relative to *b*) about the relationship between measures, but the lost information is of low relevance to the research question and is balanced by increased focus. Use of standardized coefficients might also be appropriate in certain experimental situations. Consider a hypothesis that two variables are monotonically related: that is as one variable increases the other always increases (or decreases). This is equivalent to predicting that the ranks of two variables are perfectly correlated. The hypothesis might therefore be most clearly reported using *r* for the ranks (or the equivalent Spearman correlation coefficient, r_s, for the raw scores).

A similar case can be made for r^2. For most psychological theories, explaining 100% of the variance of a phenomenon is not a realistic goal of research and the proportion of variance explained may have little theoretical relevance (Fichman, 1999). Most psychological phenomena are multifactorial – also limiting the contribution of any single predictor (O'Grady, 1982). Yet in specific situations, variance explained is a very useful tool. One particularly useful application is in contrast analysis. If an ANOVA effect has a single degree of freedom (*df*) then it is relatively simple to interpret. For effects with more than 1 *df* a single *df* contrast (in particular when defined *a priori*) is useful for interpreting the effect (Loftus, 2001). Thus a main effect could be decomposed into variance accounted for by a linear contrast and that left over, and the variance explained by the contrast could then be expressed as a proportion of the main effect: $SS_{\text{linear}}/SS_{\text{effect}}$.

The above arguments may be extended to other non-arbitrary values (e.g. values derived by theory), though these may be rare or (like $r = 0$) be more-or-less interchangeable with salient simple effect size values. Such non-arbitrary values are also potentially useful as 'absolute' benchmarks for interpreting effects. Statements about the absolute magnitude of an effect are difficult to justify under normal circumstances. This is particularly true if the absolute magnitude is related to verbal labels such as 'small', 'medium' and 'large' (e.g. Cohen, 1988). Although such 'canned' effect sizes (Lenth, 2001) are often used there is increasing consensus that they are highly misleading and should be avoided (e.g. Baguley, 2004; Glass, McGaw, & Smith, 1981). Indeed, comments about effect size that incorporate such verbal labels can lead people to misinterpret statistical findings (Robinson, Whittaker, Williams, & Beretevas, 2003). On the other hand, statements about the relative size of effect will often place the observed magnitude of the effect into an appropriate context.

Researchers often wish to compare effects obtained with non-identical measures. In these cases a transformation to a common metric is essential. In cases where the measures are re-expressions of one another (Bond, Wiitala, & Richard, 2003) this transformation can be achieved using simple effect size. Where the scales are not mere re-expressions standardization is often advocated. However, non-standardized alternatives exist that may be better suited to the problem at hand. For example, Cohen and colleagues have argued that POMP (percentage of maximum possible score)

may be superior to standardized units (Cohen, Cohen, Aiken, & West, 1999). Nor do all variables in an analysis need to be standardized (Kim & Ferree, 1981).

Where appropriate, non-identical measures may be expressed on a common scale pertinent to the goals of the research (e.g. financial cost). In cases where no suitable alternative to standardization is available, particular care needs to be taken to ensure that issues such as reliability and study design are addressed. It should never be assumed that the mere act of adopting a (superficially) standard metric makes comparisons legitimate (Morris & DeShon, 2002; Bond *et al.*, 2003). It should also be remembered that there is nothing magical about the standardization process: it will not create meaningful comparisons when the original units are themselves not meaningful (Tukey, 1969).

One putative defense of a standardized effect size metric is that it allows the comparison of effects with more than 1 *df* (e.g. η^2 or ω^2 in ANOVA). The utility of such comparisons is doubtful and reports of multiple *df* effects are generally considered less meaningful than 1 *df* effects (e.g. APA, 2001, p. 26). Multiple *df* effects are rarely replicated exactly (e.g. if the treatment represents time intervals these will rarely be identical between studies). Even if the levels of the effect were identical this would not imply identical effects. Two studies may report similar generalized ω^2 yet have radically different patterns of effects. Even so, one might reasonably use a standardized effect size statistic to test a substantive hypothesis such as $\omega^2 = 1$ (especially where separate 1 *df* effects would result in unacceptably small sample sizes). As a rule, reports of effect size should focus on 1 *df* effects (Wilkinson & APA Task Force on Statistical Inference, 1999).

Point estimates or confidence intervals?

Standard practice in psychology, if effect size is reported at all, is to report point estimates of the effect size. A superior approach is to report a confidence interval (CI). A CI conveys more information than a simple point estimate and gives an indication of a plausible range of values that the 'true' effect might take (Loftus, 2001). This use of a CI as an informal aid to interpretation of an effect is distinct from formal inference (such as a substitute for a significance test). The point estimate of an effect is easily misinterpreted because it carries no information about the uncertainty of the estimate. Imagine that a study reports a correlation of $r_{29} = .064$ and the researcher concludes that the observed relationship is negligible (or worse still that there is no relationship whatsoever). Reporting the correlation as a CI would offer protection against this incautious interpretation. An approximate CI for the effect ($-.30$, $.41$) can be obtained using the Fisher z transformation. An informal interpretation of this finding is that the correlation might plausibly be as large as .41 in the same direction (or .30 in the opposite direction). It also suggests that the study was underpowered.

The argument for reporting a CI applies equally to simple effect size. An effect with small variability is probably of more practical importance than one with large variability. Presenting simple effect size as a CI allows psychologists to consider the point estimate of an effect alongside an indication of how variable the effect is. A further advantage of CIs for simple effect size is that there is a clear distinction between the magnitude and variability of an effect (useful quantities that are confounded in standardized effect size).

Conclusions

There are strong arguments for reporting effect size in psychological research. In spite of these arguments, reporting of effect size in published work is patchy, though it may

be improving (Cumming *et al.*, 2007). One reason for this may be that researchers are uncertain of what effect size metric to report and how best to report it. There is, at present, no consensus on these issues. For example, although The APA Task Force on Statistical Inference expressed a preference for reporting simple, unstandardized effect size and use of confidence intervals, APA publication guidelines are often interpreted as encouraging point estimates of standardized effect size (Fidler, 2002). It is also likely that no single effect size metric would be appropriate for gauging the importance of an effect, comparison between findings or the diverse requirements of different forms of secondary analysis.

It is possible, however, to set out guidelines for what to report and how to report it. The main guidelines can be summarized as follows:

(1) Prefer simple effect size to standardized effect size
(2) Avoid reporting effect sizes for multiple *df* effects
(3) Prefer confidence intervals to point estimates of effect size
(4) Always include adequate descriptive statistics (e.g. sufficient statistics)
(5) Comment on the relative rather than the absolute magnitude of effects
(6) Avoid using 'canned' effect sizes to interpret an effect (Lenth, 2001)
(7) Prefer corrected to uncorrected standardized effect size estimates

Some common queries relating to some of these guidelines can be anticipated. In particular, why not routinely report both simple and standardized effect size? First, in many cases, standardized effect size can obscure the theoretical and practical importance of an effect. Researchers and consumers of research often assume that 'standardization' automatically makes a comparison meaningful (Tukey, 1969). Second, correcting these deficiencies and anticipating the range of applications research consumers will use effect size estimates for require a great deal of additional work, often for little or no gain. DeShon (2003, p. 398) notes that 'unless great care is used when correcting for measurement error, it is quite likely to make interpretation of correlation coefficients more difficult after the correction than before the correction was applied'. Similarly, Hunter *et al.* (2006, p. 594) state 'corrections for both direct and indirect range restriction are more complicated than is generally realized and are often erroneously applied'. Correcting standardized effect size also increases the width of its CI, although this should not be taken as an argument for not making the correction (see Schmidt & Hunter, 1999). Adjusting for differences in design is possible in principle, but will often be impractical without commonly agreed reference points in terms of both design and sample characteristics.

Only rarely will uncorrected standardized effect size be more useful than simple effect size. It is usually far better to report simple effect size along with descriptive statistics that allow others to derive a range of alternative effect size metrics (e.g. for comparison between studies, power calculations or meta-analysis). Reporting standardized effect size adjusted for reliability, range restriction and study design is a useful complement to reporting simple effect size. Researchers need to decide whether the additional work (and potential pitfalls) of making these adjustments is worth the effort it requires. This trade-off will be different for an individual study than for a meta-analytic review.

These guidelines are intended to sharpen routine practice in reporting effect size. This practice should be informed by the goals of the researcher and the needs of the wider research community. These guidelines are not intended to be absolute rules

and several exceptions have been explicitly considered in the preceding discussion. In determining the appropriate way to report the magnitude of an effect there is no substitute for careful thought and reflection (Tukey, 1969).

Acknowledgements

I would like to thank Koen Lamberts and two anonymous reviewers for their help in improving the clarity and focus of this paper. Thanks are also due to Dan Wright, Peter Morris, James Stiller, Raph Gillett, Jamie Murphy, Mark Lansdale, Mark Shevlin and Mark Torrance for their comments on earlier drafts.

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, D.C: American Psychological Association.

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*, 73–80.

Baguley, T., & Payne, S. J. (2000). Given-new versus new-given? An analysis of reading times for spatial descriptions. In S. Ó Nualláin (Ed.), *Spatial cognition: Foundations and applications* (pp. 317–328). Amsterdam: John Benjamins.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods*, *4*, 46–61.

Bond, C. F., Jr, Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.

Buchner, A., & Mayr, S. (2004). Auditory negative priming in younger and older adults. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *57A*, 769–787.

Clark, H. H. (1973). The language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

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

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., *et al.* (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230–232.

DeShon, R. P. (2003). A generalizability perspective on measurement error corrections in validity generalization. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 365–402). Erlbaum: Mahwah, NJ.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177.

Fichman, M. (1999). Variance explained: Why size doesn't (always) matter. *Research in Organizational Behavior*, *21*, 295–331.

Fidler, F. (2002). The fifth edition of the APA publication manual: Why its statistics recommendations are so controversial. *Educational and Psychological Measurement*, *62*, 749.

Greenland, S., Schlesselman, J. J., & Criqui, M. H. (1986). The fallacy of employing standardized regression-coefficients and correlations as measures of effect. *American Journal of Epidemiology*, *123*, 203–208.

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

Frick, R. W. (1999). Defending the statistical status quo. *Theory and Psychology*, *9*, 183–189.

Ghiselli, E. E. (1964). *Theory of psychological measurement*. New York, NY: McGraw-Hill.

Gillett, R. (2003). The metric comparability of meta-analytic effect-size estimators from factorial designs. *Psychological Methods*, *8*, 419–433.

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

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications for direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*, 594–612.

Kim, J. O., & Ferree, G. D. (1981). Standardization in causal analysis. *Sociological Methods and Research*, *10*, 187–210.

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*, 187–193.

Loftus, G. R. (2001). Analysis, interpretation, and visual presentation of experimental data. In J. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4, pp. 339–390). Methodology in experimental psychology, New York, NY: Wiley.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods*, *11*, 386–401.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, *92*, 766–777.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241–286.

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.

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*, 178–192.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with The language-as-fixed-effect fallacy: Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods*, *9*, 99–112.

Robinson, D. H., Whittaker, T., Williams, N., & Beretvas, S. N. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *Journal of Experimental Education*, *72*, 51–64.

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

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*, 183–198.

Tukey, J. W. (1969). Analyzing data: Sanctification or detective work? *American Psychologist*, *24*, 83–91.

Walter, S. D. (2000). Choice of effect measure for epidemiological data. *Journal of Clinical Epidemiology*, *53*, 931–939.

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

Wright, D. B. (2006). The art of statistics: A survey of modern statistics. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 879–901). Mahwah, NJ: Erlbaum.

Received 1 May 2007; revised version received 26 August 2008