Why are effect sizes still neglected?

Peter E. Morris and **Catherine O. Fritz** argue that it is worth describing and interpreting the size of effects as well as their significance

Psychology research is still dominated by null hypothesis significance testing (NHST). However, almost any effect will be significant if a very large sample is tested – something that has become easier to accomplish with the internet and other modern technology. Conversely, it is sometimes only possible to test a small number of participants from, for example, a special population. In that case, even a substantial difference may not be significant. NHST is dependent upon the size of the sample that is tested and the result gives only a binary 'Yes/No' decision.

Normally, our real interest is in the size of the effect. With a very large sample we might find a statistically significant effect but decide that it was too small for our further attention. Conversely, a large effect from a special, small population, even though not significant, would encourage us to seek ways of testing more participants and combining the data.

Therefore, NHST needs supplementing by informative ways of describing the effects we find. Knowing whether or not a pattern in data is significant is useful, but the data can tell us much more (as satirised by Cohen in his 1994 paper on the limitations of NHST, titled 'The earth is round (p < .05)').

In this article, we offer brief discussion of the two most commonly reported effect-size estimates: partial eta squared (η_p^2) – used with analysis of variance (ANOVA) to describe the proportion of variability associated with an effect – and Cohen's d – the difference between means of two datasets, standardised with the pooled standard deviation. Following Cohen's (1988) suggestions for interpreting effect sizes, we describe an empirically based approach to interpreting effect sizes based on the cumulative distributions of reported effect-size estimates (see Morris & Fritz, 2013). We propose interpreting effect sizes in terms of their size relative to others reported in the same general (e.g. memory) or specific (e.g. testing effect) areas of psychology. We hope that the approach we suggest will at least stimulate discussion about interpreting effect sizes and at best, provide a template for interpreting effect sizes in other areas.

Long overdue

For more than a decade most psychological journals have required the reporting of effect sizes. For example, the instructions to authors for journals published by the British Psychological Society (BPS) have included the statement that in normal circumstances effect sizes should be incorporated. The American Psychological Association (APA) *Publication Manual*, both fifth and sixth editions (2001, 2010), state: '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' (2010, p.34). And last year, the Psychonomic Society organised a special session on Improving the Quality of Psychological Science (see tinyurl.com/oaunj7f), including papers on issues relating to NHST, confidence intervals (CIs), Bayesian analyses, the importance of replications, evidence of 'p-hacking' in psychologists' research, and our own contribution on reporting and interpreting effect-size estimates (Fritz & Morris, 2012).

Effect-size estimates are useful descriptive statistics that indicate the size of the observed effects while being independent of the size of the research sample. Effect sizes are also used in calculating the power of tests to determine, for example, the number of participants required to ensure a reasonable probability of detecting an effect when one actually exists. Effect sizes also allow results to be compared and combined across many studies in meta-analyses. In Fritz et al. (2012) we describe several effect-size estimates for various types of data and analyses. Excellent guidance and discussion of effect sizes can also be found in Cumming (2012), Ellis (2010) and Grissom and Kim (2005, 2012), the latter authors being recommend by the *APA Publication Manual* (2010).

Where the units of measurement are interval or ratio scales that are very well understood by the reader (e.g. weight gain) or standardised tests (e.g. IQ tests), then simple effect sizes such as differences between means may be sufficient (Baguley, 2009). Even so, these simple effects should be supplemented by CIs, or other measures that indicate the precision of the estimates. Grissom and Kim (2012) recommend reporting both standardised and unstandardised effect sizes in such circumstances. However, psychological research often uses measurements that are specific to the materials and measures used in that research. For such data standardised effect-size estimates allow readers to better understand and compare the size of effects.

The reason for standardising the

references

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn). New York: Academic Press.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49*, 997–1003.

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

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

Fritz, C.O., Morris, P.E. & Richler, J.J. (2012). Effect size estimates: Current use, calculations and interpretation *Journal of Experimental Psychology: General, 141*, 2-18.

Fritz, C.O. & Morris, P.E. (2012, November). *Reporting beyond significance*. Invited paper as part of the Special Session: Improving the Quality of Psychological Science presented at the 53rd Annual Meeting of the Psychonomic Society, Minneapolis, MN. (Video available at tinyurl.com/oaunj7f)

Grissom, R.J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology, 79*, 314–316.

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

effect-size estimate is that most psychology studies differ from others in many ways. These include the design of the research, the stimulus material chosen (type, number of items, etc.), the way the materials are presented, the characteristics of the participants, the scoring scheme, timing considerations, and so on. For such research, simple effect-size estimates are not comparable and could be misleading. Standardising the differences goes some way towards compensating for these differences, as does the careful use of statistics such as generalised partial eta squared (e.g. Bakeman, 2005; Olejnik & Algina, 2003) that permit the removal of at least some confounding factors. Even so, as with all statistics, the values obtained need to be treated with an awareness of the situations in which they were obtained. Just as the correlations and other statistics that are so routinely used in psychological research are influenced by the same sorts of conditions just described, it is necessary to interpret and use effect sizes taking account of the context in which they are obtained. We caution against making statements such as 'the effect size for X is Y' without ensuring that the audience is aware of the conditions under which this effect size was observed. Nevertheless, effect-size estimates focus upon the observed differences rather than the 'Yes/No' of NHST (or worse still, the comparison of *p*-values) and can, with due consideration of their origins, be used to guide future research.

Yet despite publishers' requirements to include effect sizes, and exhortations to do so by the authors of statistical texts, effect-size estimates are omitted from many papers. We reported at the 2010 BPS Cognitive Psychology Section conference (Morris & Fritz, 2010) that more than half of the papers from three leading cognitive journals (the 2009 volumes of the *Journal of Experimental Psychology: Learning, Memory and Cognition; Memory & Cognition*; and the *Quarterly Journal of Experimental Psychology*) did not report effect sizes. We were asked by the incoming editor of the *Journal of Experimental Psychology: General* (JEPG) to carry out a similar review of 2010–2011 publications in JEPG and to provide a straightforward guide to authors to help understanding, selecting, calculating, and interpreting effect sizes for many types of data (Fritz et al., 2012). Again we found that effect sizes were often omitted from JEPG papers and were rarely discussed. However, in the *Journal of Experimental Psychology: Applied* and *Applied Cognitive Psychology* for 2010, 93 per cent and 72 per cent respectively of the papers reported effect sizes (Morris & Fritz, 2011). Clearly, applied psychologists have incorporated reporting effect sizes to a degree yet to be achieved in less applied research. Figure 1 summarises our surveys of the effect-size estimates just described and identifies those more commonly reported.

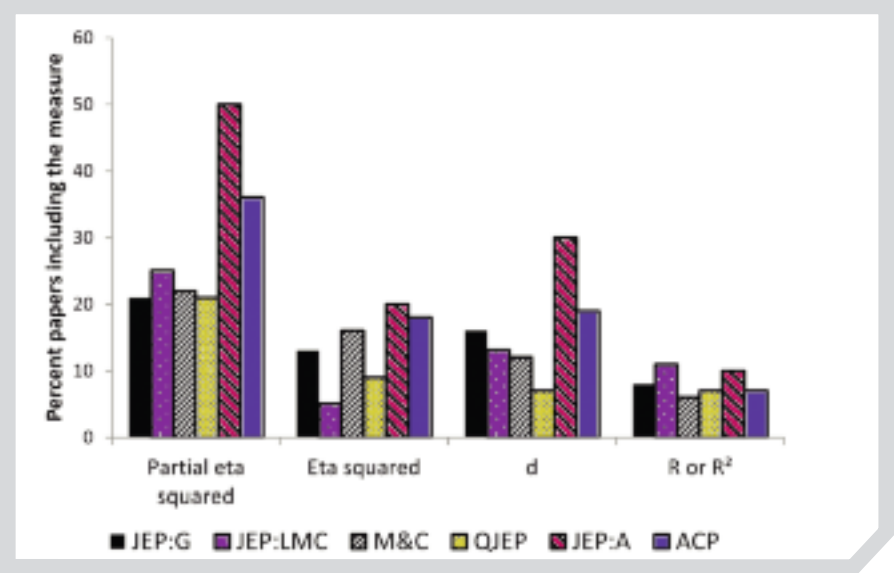

Figure 1. The percentages of papers in the *Journal of Experimental Psychology: General* (JEP:G), the *Journal of Experimental Psychology: Learning, Memory and Cognition* (JEP:LMC), *Memory and Cognition* (M&C), *Quarterly Journal of Experimental Psychology* (QJEP), *Journal of Experimental Psychology: Applied* (JEP:A) and *Applied Cognitive Psychology* (ACP) that reported the more common effect-size estimates

Why are effect-size estimates often not calculated and very rarely discussed? We have sympathy with busy researchers who report the statistics they know rather than taking the time to come to grips with statistics that they were not taught in their psychology training. We wonder, even today, how much teaching about effect-size estimation is included in statistics courses that are already hard pressed to cover NHST techniques? Statistics textbooks do introduce some effect sizes, but are a poor source for advice on interpreting even the very common ones such as eta or partial eta squared. However, reporting effect-size estimates provides readers with more useful information than the results of significance tests alone.

Some options

The most commonly reported statistic in

Grissom, R.J. & Kim, J.J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd edn). New York: Routledge

Howell, D.C. (2008). *Statistical methods for psychology* (7th edn). Pacific Grove, CA: Duxbury.

Howitt, D. & Cramer, D. (2011). *Introduction to statistics in psychology* (5th edn) Harlow: Pearson.

Morris, P.E. & Fritz, C.O. (2010, September). *From Cinderella to princess: What has and could happen to the reporting of effect sizes in cognitive publications?* Paper presented at the Annual Conference of the Cognitive Section of the British Psychological Society, University of Wales, Cardiff.

Morris, P.E. & Fritz, C.O. (2011, June). *Effect sizes and applied research.* Poster presented at the 9th Conference of the Society for Applied Research in Memory and Cognition, New York.

Morris, P.E. & Fritz, C.O. (2013). Effect sizes in memory research. *Memory.* Advance online publication. doi:10.1080/09658211.2013.763984

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.

Loftus and Palmer looked at the effect on memory of changing a single word in the questions asked of eyewitnesses of road traffic accidents

our survey was partial eta squared (η_p^2). Partial eta squared describes a proportion of variability in a sample associated with an independent variable; it is calculated as the ratio between the sum of squares for a particular factor in an ANOVA and that sum of squares combined with the sum of squares for its specific error term. It can also be calculated from the details of an *F* ratio test (e.g. Fritz et al., 2012). A key reason for the popularity of η_p^2 is almost certainly that it is offered by SPSS, so for many researchers it is readily available to report without any calculations or careful consideration of what the statistic estimates.

Clearly η_p^2 deserves attention and consideration because of its widespread use. Nevertheless, it is worth noting that as an effect-size estimate, it has limitations: Partial eta squared is rarely mentioned in statistical texts for psychologists (e.g. Howell, 2008; Howitt & Cramer, 2011) or in specialist texts on effect-size statistics (e.g. Cumming, 2012; Ellis, 2010; Grissom & Kim, 2012). When it is mentioned, it is almost always to criticise it in comparison with the authors' preferred effect-size statistics. One should remember that η_p^2 is a sample statistic and therefore will overestimate the size of the effect in the population, especially when the sample sizes are small. However, this criticism, although technically correct, also applies to the correlations (*r*) and multiple correlations (*R*) that are routinely reported by psychology researchers. Methods for correcting the overestimation for correlations have existed almost as long as correlations have been used, but psychology researchers usually rely on the uncorrected correlations. In a similar way, researchers could acknowledge the possibility of overestimation by η_p^2, especially with small samples. In practice, the overestimation of η_p^2 is quite small when the sample sizes are moderate to large and there are few levels of the independent variables (IVs). For example, the difference between η_p^2 and its population estimate counterpart, partial omega squared (ω_p^2), is less than .03 and is usually less than .02 when the total number of participants is 50 or greater and there are four or fewer levels of the IV.

Alternatives to η_p^2 that address some of these problems are available. Omega squared (ω^2) is the most popular effect-size estimate among textbook writers, and we agree with the respectability of this estimate of the proportion of variability in the population. However, we found that ω^2 was almost never reported. There are three likely reasons for this neglect: (1) commonly used statistical software such as SPSS does not calculate ω^2, (2) there are difficulties in applying ω^2 to repeated measures designs, and (3) ω^2 values are usually smaller than η_p^2 values, especially when sample sizes are small.

Generalised eta squared (η_G^2) is another effect-size estimate related to η^2 and η_p^2; it is especially useful when making comparisons across studies (Olejnik & Algina, 2003). In Fritz et al. (2012) we describe in some detail the calculation of η_G^2 which we believe to be a more useful estimate than η^2 and η_p^2, and one which can be used with repeated measures (Bakeman, 2005). Nevertheless, it, like η_p^2, is a sample statistic and so can overestimate population effect sizes when samples are small.

Cohen's *d* is generally more respected than η_p^2 as an effect-size estimate by statistical text authors; extended discussions will be found in several recent texts (e.g. Cumming, 2012; Grissom & Kim, 2012). It is easy to calculate and there are several free online calculators for *d* (see e.g. tinyurl.com/kbu3m73). Even so, we found many occasions when *d* was not reported although it would have been appropriate. Often *d* was reported in conjunction with *t* tests, but not with ANOVA, even where factors had only two levels. This omission is unfortunate, because it is common to be more interested in comparing means within a factor than in the overall influence of the factor itself. It would often be very useful if the comparisons between means were reported in addition to the overall effect size. We almost never observed the reporting of more than one effect size for a particular set of data, but often doing so would assist the interpretation of the observed effects (see Fritz et al., 2012, for further discussion).

Helping to evaluate findings

Researchers who report effect-size estimates usually report η_p^2 or *d*, but do not interpret the observed effect size beyond a brief reference to Cohen's (1988) guidelines for small, medium and large. We believe that one of the main reasons why effect sizes are often omitted and, when included, are rarely discussed, is that researchers have no ready source to guide interpretation of their values. Recently (Morris & Fritz, 2013), we have addressed the interpretation of η_p^2 and *d*

by following Cohen's (1988) suggestion that the values should be considered in the context of research in an area. Within our own main research area of memory we tabulated reported values of η_p^2 and *d* in memory publications in 2010 and supplemented these values with calculated values where possible. From nearly 3000 values for η_p^2 and 720 values for *d*, we produced tables of cumulative frequencies for η_p^2 (enabling easy look-up of *cumulative percentage partial eta squared* values: $C\eta_p^2$) and cumulative frequencies for *d* (for *cumulative percentage d* values: *Cd*). (See Morris & Fritz, 2013, for these tables.). Thus, if a memory researcher observes η_p^2 = .34 in their latest study they will find a value of $C\eta_p^2$ = 69.1 per cent when they consult the table. In other words, their observed effect size is larger than 69 per cent of other η_p^2 values in recently published memory research.

Our aim in producing the cumulative effect-size tables is to help researchers evaluate their η_p^2 and *d* findings against other published research in their research domain. Until now, the only available guides were those for small, medium and large effects offered by Cohen (1988) based, by his own admission, on his intuition. At least for memory and applied research Cohen's guideline values for η_p^2 (.01, .06 and .14) are substantially smaller than the observed medians and quartiles. For memory research we found quartile and median values of .08, .18 and .41, with very similar results for applied research. Similar surveys for other research areas could provide valuable information to help researchers evaluate their effects.

For *d*, Morris and Fritz (2013) found that the cumulative percentage (*Cd*) median and quartile values (.25, .57, .99) were similar to those suggested by Cohen for small, medium and large effects (0.2, 0.5, 0.8). There are also other useful translations to aid interpretation of *d* that we describe in Fritz et al. (2012). These are the probability of superiority (*PS*; Grissom, 1994; Grissom & Kim, 2012) and the percentage of non-overlap of the distributions (U_1: Cohen, 1988). *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 is the percentage of non-overlap between the two distributions.

As an example, we have taken a classic paper (Loftus & Palmer, 1974) and supplemented one of its findings with information about its effect size. Loftus and Palmer were interested in the effect of changing a single word in the questions asked of eyewitnesses. The crucial question asked participants how fast the cars were going when they hit/smashed into each other. Loftus and Palmer reported that the mean speed estimates were: 'hit' 8.00 mph and 'smashed into' 10.46 mph, and that the difference was statistically significant, $t(98) = 2.00$, $p < .05$. So, it was significant, but did that matter? Although Loftus and Palmer gave no values for the variability of their data, it is possible to calculate the value of *d* from their *t* test (Fritz et al., 2012): $d = 0.4$. Using the tables from Morris and Fritz (2013), a *d* of 0.4 has a $Cd = 37.6$. That is, the *d* found by Loftus and Palmer in this particular condition is bigger than over 37 per cent of values of *d* in recent memory research. An impressive influence from the change of one word. Furthermore, from the table in Fritz et al. (2012) we can see that the *PS* for a *d* of 0.4 is 61. That means that if paired samples were drawn at random, one from each of the sets of data for participants who heard 'hit' or 'smashed into', there would be a higher estimate of the speed for the 'smashed into' participants on 61 per cent of occasions, and higher estimates for the 'hit' participants in 39 per cent of cases. Also, from the table in Fritz et al. we can see *that for* d = 0.4 the value of Cohen's U_1 is 27, meaning that the 27 per cent of the distributions for 'hit' and 'smashed into' would not overlap. This information about the size of the effect helps researchers and readers to better conceptualise and interpret it.

Alternatives and confidence intervals

So far, we have discussed just η_p^2 and *d*. Interested readers should consult Fritz et al. (2012), Ellis (2010), Cumming (2012) or Grissom and Kim (2012) for introductions to other useful effect-size estimates. Some authors favour the point-biserial correlation (r_{pb}) instead of *d* (see, e.g. Grissom and Kim, 2012 for a discussion), but we found this statistic very rarely reported. The values for *r* in Figure 1 are almost always for correlations that were reported as part of the general analysis and not specifically as effect-size estimates.

There are effect-size statistics appropriate to most types of data, including categorical (e.g. phi, Cramer's *V*, and lambda) and ordinal data in Mann–Whitney and Wilcoxon nonparametric tests (*z*). Regression analyses are typically accompanied by R^2 and adjusted R^2. We suspect that the diversity of effect-size statistics creates some confusion and contributes to their neglect. For some data analyses people may have difficulty in deciding which statistics to use and how to calculate them; the Fritz et al. (2012) paper provides some easily accessible help in this regard. Fortunately, most of the statistics are relatively simple and are easy to calculate. In many cases, it is also possible to convert one statistic into another, and we provide formulae in Fritz et al.

The *APA Publication Manual* says that confidence intervals for effect sizes should be included when possible. This advice is sensible because an effect-size estimate is a point estimate based on a sample and the likely range within which a replication might fall is given by CIs. However, we found very few publications reported CIs for effect sizes. One reason may be that CIs for effect-size estimates such as *d* are asymmetric and require estimating through successive approximations. However, Cumming (2012) has provided excellent software for calculating CIs for *d* (tinyurl.com/mtrnlhl). Where authors have reported CIs for effect sizes they have invariably used the 95 per cent interval. Although 95 per cent CIs are an appropriate replacement for NHST, and can be used to do so with effect-size estimates, they seem unnecessarily broad as guides to the likely range of an effect size. We recommend also reporting 68 per cent CIs for effect-size estimates (Fritz & Morris, 2012); these encompass over two thirds of the likely values for the population parameter, but their range that they cover is half of that for the 95 per cent CI.

Conclusion

Our hope is that researchers will report and interpret effect-size estimates with a similar emphasis and enthusiasm to that associated with statistical significance. As the practice becomes more widespread, we predict that researchers will recognise the contribution that effect-size estimates can make to the interpretation of data and our understanding of psychological processes.

Peter E. Morris *is Honorary Visiting Professor of Psychology at the University of Northampton, and Emeritus Professor at the University of Lancaster*
p.morris@lancaster.ac.uk

Catherine O. Fritz *is at the University of Northampton*
catherine.fritz@northampton.ac.uk