THE INTERPRETATION OF EFFECT SIZE IN PUBLISHED ARTICLES

Rink Hoekstra
University of Groningen, The Netherlands
R.Hoekstra@rug.nl

Significance testing has been criticized, among others, for encouraging researchers to focus on whether or not an effect exists, rather than on the size of an effect. Confidence intervals (CIs), on the other hand, are expected to encourage researchers to focus more on effect size, since CIs combine inference and effect size. Although the importance of focusing on effect size seems undisputed, little is known about how often effect sizes are actually interpreted in published articles. The present paper will present a study on this issue. Interpretations of effect size, if they are presented in the first place, are categorized as unstandardized (content-related) or standardized (not content-related). Moreover, the interpretations of effect size for articles that include a CI will be contrasted with articles in which significance testing is the only used inferential measure. Implications for the current research practice are discussed.

INTRODUCTION

The importance of reporting effect size in a scientific article seems undisputed. Presenting a measure of effect size is considered crucial for a proper understanding of the meaning of the data. In the 6th edition of the APA Manual (2009) it is stated that "[f]or the reader to appreciate the magnitude or importance of a study's finding, it is almost always necessary to include some measure of effect size in the Results section" (p.34). Many have argued that, rather than focusing on significance test outcomes, the focus should be on effect size instead (e.g., Cohen, 1990). Indeed, not only has been argued that effect sizes measures should be reported, they should also be interpreted (e.g., Cumming, 2012; Rosnow & Rosenthal, 2009).

There are several ways to define effect size. Rosnow and Rosenthal (2009) consider it "the magnitude of a study outcome or research finding, such as the strength of the relationship obtained between an independent variable and a dependent variable" (p. 6). This includes a simple difference of two means, which can be considered the simplest way to present effect size (e.g., Riopelle, 2000; Cohen, 1988). In these cases the effect size is measured on the scale of interest. In some cases, however, it is harder to interpret these outcomes (for example because the scale on which is measured is not known to the reader, and in those cases some (e.g., Cohen, 1988; Rosnow & Rosenthal) recommend a measure that is independent of the scale which is measured on. Cohen proposed to standardize the raw difference between the two sample means, resulting in a value referred to as *d*, Cohen's *d*, or δ. Related alternatives are Hedge's *g* (Hedges, 1982), which uses the pooled standard deviation, and Glass' Δ (Glass, McGaw, & Smith, 1981), which uses the standard deviation of the control group only. In the present study, both standardized and unstandardized values are considered measures of effect size. This is also how effect size is defined in the APA manual (2009): "Effect sizes may be expressed in the original units….. but also in some standardized or units-free unit (e.g., as a Cohen's *d* value)" (p. 34). Some (e.g., Kirk, 1996), however, have a narrower definition, and use the term exclusively for the standardized version. Kirk uses *effect magnitude* for the coordinated term.

The interpretation of an effect size is not straightforward. In case of unstandardized measures, its interpretation largely depends on the scale on which it is measured. Although there may exist "objective" criteria for interpretation in these cases, more often than not the interpretation also depends on the person interpreting the outcomes. Although this is not problematic per se, authors might feel reluctant to give an evaluation of the outcome that can be considered subjective. In case of standardized measures of effect size, however, more guidelines exist on how to interpret these outcomes. Cohen (1988) suggested values that could be interpreted as "small", "medium" or "large" (for *d* theses were 0.2, 0.5 and 0.8, respectively, and he also provided guidelines for seven other effect size measures), but he also explicitly noted that a sound interpretation was content-depended, and should not rely on arbitrary rules. That is, even in case of standardized effect sizes, the interpretation is not trivial per se.

In K. Makar, B. de Sousa, & R. Gould (Eds.), *Sustainability in statistics education. Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9, July, 2014), Flagstaff, Arizona, USA.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

Null-hypothesis significance testing (NHST), the most frequently used technique in the social sciences, is often criticized for drawing the attention away from the size of the effect (e.g., Cumming, 2012; Kirk, 1996), and for drawing the focus solely on whether or not the null hypothesis is true instead. As is well known, significance (or the absence of it) by itself does not tell you anything about the size of an effect, which has lead some to make a distinction between statistical significance and practical or clinical significance. Given that researchers typically start from a content-related research question, one could expect them to be at least as interested in practical as in statistical significance, but, given that statistical significance testing is often seen as a criterion for publication, the focus in the current research practice seems to be on statistical significance. Sohlberg and Andersson (2005), however, believe that it is getting easier for those who weigh in effect sizes in addition to *p*-values to get their article published compared to those who only base their conclusions on NHST, but they do not refer to evidence to substantiate their claim.

In the discussion about the lack of attention for effect size when researchers use NHST, confidence intervals (CIs) have often been proposed as an alternative. Schmidt (1996), for example, argues that "[i]n our graduate programs we must teach that for analysis of data from individual studies, the appropriate statistics are point estimates of effect sizes and CIs around these point estimates" (p. 116). Velicer et al. (2008) argue that effect sizes measures have become the basis of power and meta-analysis, and that they can be used to make predictions on the basis of a theoretical model. CIs, according to them, can subsequently be used to express how strongly the theory is supported by the data. With CIs, effect size is explicitly shown (actually, the CI is constructed around the effect size). Some even define the range of the interval as the range of plausible effect sizes (e.g., Cumming, 2012), but it can be argued that this is not what can soundly be concluded from a CI (e.g., Hoekstra, Morey, Rouder & Wagenmakers, 2014), given its frequentist nature. Because CIs are said to draw the focus away from the null-hypothesis (for CIs, no null-hypothesis is needed), it can be expected that effect size would be mentioned in the interpretations more often. Hoekstra, Finch, Johnson and Kiers (2006) showed that in almost all articles they studied, some measure of effect size was reported. They did, however, not study whether the size of the effect was *interpreted*.

In an experimental study by Hoekstra, Johnson and Kiers (2012), thirty researchers were asked to write down an interpretation of outcomes that were either presented by means of CIs or by NHST, including an effect size. It was found that participants referred somewhat more frequently (57% versus 42%) to the size of the effect when the data were presented by means of CIs compared to NHST. Although this confirms the expectation that CIs draw researchers' attention more to the size of the effect, it is hard to generalize to generalize these data to how outcomes are interpreted in articles. Cumming (2012) provides a few "good practice" examples of how, according to him, a proper interpretation of effect size could look like, but since they are selected for this purpose, it is unsure whether they are exemplary for the current practice of the interpretation of effect size in practice. Indeed, we are not aware of any study on the interpretation of effect size.

In summary, effect size measures are widely considered crucial. Nevertheless, little is known about their use in published articles. Are they merely presented, or also interpreted? And if they are interpreted: is this interpretation based on the unstandardized or the standardized measure? Furthermore, it will be studied whether in studies in which CI are presented, the effect size is more often interpreted, as many who argue in favor of the use of CIs have claimed.

METHOD

Articles

For the current study, a sample of 33 articles from the *Psychonomic Bulletin & Review* were reviewed. The journal was chosen for a few reasons: it accepts articles with a wide variety of topics within psychology, and the journal is relatively prominent, with an impact factor of 2.2 in 2012. All articles in the first two volumes of the journal in 2013 were included in the sample, provided that they included quantitative outcomes. The Result and the Discussion sections of the articles were scored by the author of this paper by means of a checklist. The introduction and method were excluded because typically they do not contain outcomes of the study at hand.

Checklist

For each article that was reviewed, a 7-item checklist was used. Items were scored "1" if at least one occurrence was found, and "0" otherwise. All items where only scored whenever main outcomes of the study were presented. That is, the proportion of removed participants would not be considered an effect size measure, unless this is one of the variables of interest of the particular study. The following items were checked for:

1) Is a measure of unstandardized effect size presented?
2) Is a measure of standardized effect size presented?
3) Is a measure of unstandardized effect size interpreted?
4) Is a measure of standardized effect size interpreted?
5) Is there any mention of an effect size in the discussion section?
6) Is there any comparison made between a found effect size and effect sizes in earlier literature?
7) Is a CI reported?

Unstandardized measures of effect size include any reference to effect size that is interpretable on the scale of interest. This includes means, difference of means, proportions, and correlations, either reported in the text, or in figures or tables. Standardized measures are those that can be interpreted independent of the scale of interest, and include Cohen's d, η^2, or r^2, in text, figure or table. An interpretation of effect size was scored whenever an effect size measure was given an evaluation related to the size of the effect. That is "this is a strikingly large effect" is considered an interpretation, whereas "there clearly was a difference between group A and group B" is not. A reference to effect size in the discussion section was scored whenever an effect size was mentioned in the text of the discussion section. CIs were counted whenever they appear in the text, in a figure or a table, provided that they were explicitly referred to as CIs. Standard error bars, for example, could given some assumptions be considered as a 68% CI, but they are not in this study. According to us, standard error bars are usually used as a descriptive measure, whereas a CI is an inferential measure, and therefore they are not treated equally, despite the fact that they are comparable mathematically speaking.

Analysis

Since the articles in our sample are published in the same journal, they cannot be considered a random sample from all peer-reviewed articles in the social sciences. For that reason, no inferential statistics are presented in this manuscript. Furthermore, the sample is still relatively small, which makes it hard to make clear inferences as well. Nevertheless, the outcomes can give an indication of how effect size is used in a prominent journal in psychology.

RESULTS

All of the 33 articles that were reviewed for this pilot study were experimental articles presenting new results. In every article, some measure of effect size was presented at least once. In thirty-one articles (93%), an unstandardized measure of effect size was given, whereas in 14 articles (42%) some standardized measure of effect size was reported. In a majority of the articles (76%), however, not a single interpretation was given to these outcomes. Moreover, few comparisons were made between the size of the effect that was found and previous effects (9%), and effect size measures were hardly mentioned in the discussion.

Table 1. Frequencies and Proportions for the Presentation and Interpretation of Effect Size

	Number of articles	Percentage of articles
Some measure of effect size (total)	33	100%
Unstandardized	31	93%
Standardized	14	42%
Interpretation of effect size (total)	8	24%
Unstandardized	4	12%
Standardized	4	12%
Effect size mentioned in discussion	5	15%
Comparison effect size earlier literature	3	9%
Confidence interval reported	4	12%

It was hypothesized that in cases that a CI was presented, effect sizes would be more often interpreted than in cases they were not. In the current sample, however, only four times at least one CI was observed in an article. Although the occurrence of an interpretation of effect size was found more often (1 out of 4 = 25%) than in the other cases (3 out of 29, 10%), these numbers are too small to warrant a conclusion on this issue.

DISCUSSION

Since effect size is widely considered crucial for a proper understanding of study outcomes, it is not surprising to find that in all articles in our sample some measure of effect size was reported for the main outcome or outcomes. Nevertheless, arguably more surprising, effect sizes measures were only interpreted in a minority of the articles, and hardly mentioned in the discussion section, or compared to effect sizes found in earlier studies. Apparently, effect size is viewed as something that needs to be reported, but not necessarily interpreted. Only judging from the relatively small sample of published articles, effect size does not seem to be considered as crucial by authors as many who argue in favor of effect size (e.g., Cohen, 1990; Cumming, 2012) would like it to be.

At least two explanations could explain the findings in our paper. The first is the continuing dominance of significance testing. Although its use and usability within the social sciences have been criticized for decades (see e.g., Kline, 2012), many researchers seem to rely on NHST as a first and main outcome measure to base their conclusions on. If NHST is seen as a measure of importance of the outcome, adding effect size might not add much to it. A second explanation could be that researchers feel uncomfortable interpreting their findings in such a way that could be considered “subjective” by others. Despite the limitations of NHST, there can be no discussion about whether a certain *p*-value is significant, given a certain significance level. By focusing on the significance only, researchers might think that they avoid difficult discussions about their interpretation of the outcomes. It would be ironic if this were true, given the number of articles that have been written about the problematic interpretations of NHST (Kline).

Of course, the number of articles that were scanned for this paper was relatively small. The main reason for that is that these results are preliminary, and part of ongoing research. A second limitation is the fact that only articles from one journal were used. Although both hamper the generalizability of the outcomes, the results already indicate quite clearly that interpreting effect sizes is far from standard in published articles in this journal, and it seems unlikely that it will be completely different in other articles and different journals. If that is indeed the case, effect size has a much less prominent place in the social sciences than one would expect given how frequently its use and interpretation has been advocated, and given that it is an outcome that summarizes that what one would expect the researcher to be interested in.

REFERENCES

American Psychological Association (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

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

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

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

Hoekstra, R., Finch, S., Kiers, H. A. L.,& Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p-values. *Psychonomic Bulletin & Review*, *13*, 1033–1037.

Hoekstra, R., Johnson, A.,& Kiers, H. A. L. (2012). Confidence intervals make a difference: Effects of showing confidence intervals on inferential reasoning. *Educational and Psychological Measurement*, *72*, 1039–1052.

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (in press). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*.

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

Kline, R. B. (2013b). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington,DC: APA Books.

Riopelle, A. J. (2000). Are effect sizes and confidence levels problems for or solutions to the null hypothesis test? *Journal of General Psychology*, *127*, 198-216.

Rosnow, R. L. & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift für Psychologie/Journal of Psychology*, *217*, 6-14. doi: 10.1027/0044-3409.217.1.6

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Sohlberg, S., & Andersson, G. (2005). Extracting a maximum of useful information from statistical research data. *Scandinavian Journal of Psychology*. *46*, 69-77.

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology: An International Review*, *57*, 589–608. doi:10.1111/j.1464- 0597.2008.00348.x