Misunderstandings and omissions in textbook accounts of effect sizes

Paul H. Morris

Abstract

There have been frequent attempts in psychology to reduce the reliance on null hypothesis significance testing (NHST) as the criterion for establishing the importance of results. Many authorities now recommend the reporting of effect sizes (ESs) as a supplement or alternative to NHST. However, there is extensive specialist literature highlighting problems associated with the use and interpretation of ESs. A review of the coverage of ESs in over 100 textbooks on statistical analysis in behavioural science revealed widespread neglect of ESs and the relevant critical issues that have widespread coverage in the more specialist literature. For example, many textbooks claim that ESs should be interpreted as a simple measure of the practical real-world importance of a result despite the fact that ESs are profoundly influenced by features of design and analysis strategy. We seek to highlight areas of misunderstanding about ESs found in the pedagogical literature in light of the more specialist literature, and make recommendations to researchers for the appropriate use and interpretation of ESs. This is critical as statistics textbooks have a crucial role in the education of researchers.

Keywords: effect size, the new statistics, NHST, textbooks, methodology, statistical reporting

There is now widespread acceptance that there has been an over reliance on null hypothesis significance testing (NHST) as a means of determining the credibility and importance of results from inferential tests in psychological research (Boring, 1919; Chandler, 1957; Cohen, 1994; Cumming, Fidler, Kalinowski & Lai, 2012; Denis, 2003; Gigerenzer, 1993; Meehl, 1978; Rozeboom, 1960). Indeed, hundreds of articles documenting the shortcomings of NHST have been published over the decades but as Rozeboom (1997) states, "It is a sociology of science wonderment that this statistical practice has remained so unresponsive to criticism" (p. 335). It is also noteworthy that there are consistent misunderstandings about the true meaning of NHST (Berger, 2003).

Critics of NHST have proposed a number of alternative approaches (e.g. Bayesian modelling), but the most common suggestion to improve statistical interpretation has been to include effect sizes (ESs) in addition to *p* values (Cohen, 1962; 1988; 1994; Vaughan & Corballis, 1969; Wilkinson & APA Task Force on Statistical Inference, 1999). The reporting and interpretation of ESs also play a critical role in what Cumming (2014) terms "The New Statistics" (p. 7). However, the calculation and interpretation of ESs is by no means straightforward and there is an extensive specialist literature covering difficulties associated with the use of ESs (Baguley, 2009; Cortina, & Landis, 2009; Fern & Monroe, 1996; Grissom & Kim, 2014; Morris & DeShon 2002; O'Grady, 1982; Olejnik & Algina, 2003; Onwuegbuzie & Levin, 2003; Osborne, 2003; Petrinovich, 1979; Richardson, 1996; Rosenthal, Rosnow & Rubin, 2000; Sackett, Laczo & Arvey, 2002). A recent article by Pek and Flora (2018) provides a guide to the appropriate reporting of ESs.

Despite widespread attempts to encourage researchers to include and interpret ESs in research papers, change in relevant practice has been very slow (Cumming et al., 2007; Sun,

Pan & Wang, 2010). Gigerenzer, Krauss and Vitouch (2004) have identified the statistics textbook as one important reason for the glacial nature of change. Indeed, there is a growing literature on the influence of the information and misinformation found in psychology textbooks (Costa & Shimp, 2011; Costall & Morris, 2015; Levine, Worboys & Taylor, 1973). For most research psychologists, statistical analysis is a tool of the trade rather than a research interest of itself. Many active researchers refer to statistics textbooks for guidance with regard to the conduct and interpretation of statistical analyses and textbooks may have an important role in determining how many researchers use and interpret ESs. Textbooks also often reflect the received wisdom about a particular topic. Therefore, a careful examination of the treatment of ESs in textbooks may be very important in understanding how research psychologists are being advised with regard to ESs. In this paper, we highlight the major issues surrounding the use of ESs covered in the specialist literature and show how the neglect of these issues is leading to consistent misunderstandings about ESs in textbooks.

We used a variety of methods to source information from textbooks, some systematic and some less systematic. We looked for relevant books (books with ES in the title; general books on research methods and statistical analysis for the behavioural sciences and related disciplines) from all major publishers (education publishers in the top 10 by income, source of information was *Publishers Weekly*), we also used "effect size" as a search term in Google Books. In our analyses below, we include evidence from over 100 books.

Are ES transparent measures of the practical "real world" importance of a result?

In this section we address the issue of whether ESs can be regarded as a measure of practical importance. There are two types of ES, standardized and unstandardized. Standardized ESs have been defined by Baguley (2009) as "a standardized measure of effect […..]which has been scaled in terms of the variability of the sample or population from

which the measure was taken. In contrast, simple effect size (Frick, 1994) is unstandardized and expressed in the original unit of analysis" (p. 604). In the overwhelming majority of journal articles and books the use of the term ES refers to standardized ES. We focus on standardized ESs in this section.

One of the criticisms of NHST is that it provides no indication of the magnitude of any effect, and therefore indicates nothing about the practical importance of a result. In contrast it seems intuitively reasonable that ESs may provide evidence for the practical importance of a result. However, there is a fundamental problem with treating ES as a measure of practical, real world importance – ESs are profoundly influenced by experimental design and the type of analysis employed (Fern & Monroe, 1996; Onwuegbuzie & Levin, 2003). For example, the use of an independent groups or repeated measures design can have a major influence on ES. In an independent groups design, individual differences variance is included in the general error term and is thus part of the calculation of the ES. In a repeated measures study, the portion of variance attributable to individual differences is typically treated as a separate effect and is thus excluded from the comparison of treatment and error variance in the calculation of the ES (Maxwell & Delaney, 1990). Therefore, repeated measures designs often produce larger ESs than independent groups designs (Keppel, 1991; O'Grady, 1982). For example, data (Levy, 1973) producing the following means and standard deviations: M = 8.00, SD = 4.85; M = 11.00, SD = 5.43; M = 14.00, SD = 3.74[1] were analysed using a one-way repeated measures analysis of variance (ANOVA) and then rearranged for analysis using a one-way independent groups ANOVA. The eta squared produced from the repeated measures ANOVA was .66 whereas the eta squared for the independent groups ANOVA was less than half at .30. There are better measures of ES than eta squared, however, these are the

[1] It is typically the case that repeated measures designs produce larger ESs as the repeated measures are often highly correlated.

results that Statistical Package for the Social Sciences (SPSS [IBM], 2017) would have produced and we suspect that many researchers would have taken the eta squared values at face value[2].

A related problem is that depending on the ES employed, the magnitude of an effect associated with a particular independent variable can change when other variables are added into the model. This is a particular issue for partial eta squared (the default ANOVA ES measure in SPSS). We produced a data set where the variance associated with one independent variable was 224.50 and the error variance was 10946.10. We added another independent variable to the analysis that accounted for a large amount of variance which reduced the size of error term from 10946.10 to 683.60. The partial eta squared for the first independent variable increased from .02 to .25 with the inclusion of the second variable. We would again stress that there are better measures of ES than partial eta squared, however, the point is that the selection and interpretation of ES statistics is by no means straightforward. The addition of levels within a variable can also change the magnitude of the ES (O'Grady, 1982). There have been attempts to correct for various aspects of design on ES, with varying success (e.g. Cooper, Hedges & Valentine, 2009; Dunlap, Cortina, Vaslow & Burke, 1996; Morris & DeShon, 2002; Olejnik & Algina, 2003).

There are other factors that can profoundly affect study ES. Range restriction and attenuation both affect the accuracy of the estimate of the true magnitude of ESs (Bobko, Roth & Bobko, 2001; Osborne, 2003; Sackett, Laczo & Arvey, 2002). Both these effects are more traditionally associated with correlational techniques but can also affect tests of difference (Bobko, Roth & Bobko, 2001). The range restriction effect is a function of sampling technique. If the sample for a test is taken from the centre of the distribution,

[2] SPSS reports partial eta squared by default but there is no difference between partial eta squared and eta squared for simple one factor designs.

ignoring the extreme tails, the sample ES is likely to be an underestimate of the true magnitude, whereas if only the extreme tails are sampled, the ES is exaggerated. Attenuation typically leads to the underestimate of the population correlation that is a function of the unreliability of the measurement of the two test variables, although attenuation can result in overestimation in more complex models. Correction formulae are available to ameliorate both effects (Ghiselli, 1964; Wiberg & Sundström, 2009). Despite the fact that these problems have a long history in psychology (Pearson, 1904; Spearman, 1904), they are widely neglected in any discussion of ESs in almost all textbooks in the context of ESs (see Baguley [2009] for a complete discussion on the limitations of standardized ESs).

We did find a small number of textbooks that described how study design directly controls ES magnitude. In fact, Cohen (1988), who has been perhaps the most influential psychologist in raising the importance of ESs, was emphatic that there are a number of issues regarding the conduct of a study (including study design) that can be a very important determinant of ES magnitude.

> "Thus, operative effect size may be increased not only by improvement in measurement and experimental technique, but also by improved designs." (Cohen, 1988, p. 13)[3]

However, in the overwhelming majority of textbooks, readers are explicitly encouraged to view ESs as a simple index of the practical importance of a result. The relationship between the type of design and type of analysis employed and ES magnitude is entirely neglected. A very simple and neat narrative is presented in many textbooks: NHST provides

[3] See also references 13, 17, 56,71 and 110 for quotes in the online supplement to this paper. All subsequent numbers in footnotes refer to supporting quotes referenced by number in the online supplement.

information about whether or not the result is due to chance, whereas ESs are described as providing information about the practical, real world significance of the result.

> "Therefore, regardless of the level of statistical significance, one can ask to what extent the result is useful for understanding how one variable practically influences the other. In fact, evaluators often refer to effect sizes as 'practical significance' over and against statistical significance." (Abbot, 2010, p. 20.)[4]

It is clear that a large number of authors make very strong claims about the interpretation of ESs. ESs are alleged to provide an unproblematic index of the practical, real world importance of the results obtained and many explicitly use both these terms. Indeed, many authors introduce ESs using a subheading of "practical significance" (e.g. Stangor, 2014; Steinberg, 2010; Stufflebeam & Shinkfield, 2007; Weiner, Schinka & Velicer, 2012). A related claim is that ESs are a (best) estimate of population or real-world differences. "Effect size […..] The degree to which a phenomenon is present in the population…." (Cohen, 1988, p. 10.)[5]

Therefore, it appears that some authors actively encourage readers to regard a study ES as not local to the study, but as a measure of the real world, population effect. In the overwhelming majority of textbooks there were no caveats about the interpretation of ESs. Similarly, there is no recognition that ES estimates need to be treated with caution as ESs are strongly influenced by decisions about design and analysis, or that ESs are not unbiased indicators of effects in the 'real world'. The fundamental problem with using ESs in the

[4] See also 1, 4, 6, 21, 24, 25, 27, 29, 30, 40, 48, 50, 53, 62, 65, 70, 72, 74, 79, 80, 89, 93, 94, 98, 102, 109, 113, 114, 121, 123, 124, 126, 128, 132.
[5] See also 12, 15, 16,28, 49, 58, 66, 73, 76, 86, 106, 127.

context of experiments is elegantly summarised by Petrinovich (1979) in a paper on functionalism.

> "The very power and elegance of the experiment render it inappropriate to determine the probable importance of an independent variable to control a dependent variable. If the experiment is done well enough, if the experimenter is good enough and is able, by direct or by statistical control, to eliminate all potentially relevant variables from exerting any influence, then the variable left free to vary must account for a large proportion of the total variance in the dependent variable, *even though it might control only a miniscule proportion of the variance in natural settings where all variables are free to covary unhindered by the experimenter.* The experiment, then, is ideal to determine possibility, but it falls short of being adequate to determine external probability." (p. 376) (italics original)

Experiments were developed to investigate causal relationships between variables; they were not developed to assess the magnitude of effect. The superiority of experiments to other forms of investigation is that they allow strong claims about causality. The causal relationships found in experiments can also exist in the "real world" environments, but it is highly unlikely that ESs will ever be a precise measure of "real world" ESs.

ESs in correlational designs, which are usually less controlled than experimental studies, may be closer to "real world" ESs. However, it is sometimes claimed that the term ES should be restricted to experiments.

> "I do not like effect size as a term, at least when it is used in studies that do not assess interventions. The problem, in my view, is that its name (effect

> size) implies that one variable in a relationship affects (causes) another." (Rosenthal, & Rosenthal, 2011, p. 80.)

The arguments presented in this paper are largely restricted to experimental studies, because this is the context in which most arguments about ESs are presented in the literature. Arguments surrounding the use of ESs in correlational studies could be very different.

In summary, there is much evidence that suggests that ESs should not be treated as simple indexes of the practical importance of a result, however, many textbook authors are arguing that the main function of ESs is to provide information about practical, real world effects.

What is a big ES and what is a small ES?

Cortina and Landis (2009) present strong arguments that small effect sizes can sometimes be important and large effect sizes can be unimportant (see also Lenth [2007] who presents a really straightforward account of how to approach this problem when performing power calculations). They warn of the dangers of equating a fixed set of ES magnitudes with the importance of the result. As an exceptionally influential statistician, Cohen (1988) may be the inadvertent originator of the problem that Cortina and Landis raise, as Cohen does indeed suggest that certain ES magnitudes should be considered as small, medium and large and many textbooks adopt such an interpretation. The implicit assumption is that a small ES means that the result is unimportant whereas a large ES is always important.

> "Luckily, Cohen (1988) has made some widely accepted suggestions about what constitutes a large or small effect: $r = 0.10$ (small effect) […..] $r = .30$ (medium effect […..] $r = .50$ (large effect) […..]. We can use these

> guidelines to assess the importance of our experimental effects (regardless of the significance of the test statistic)." (Field & Hole, 2002, p. 153.)[6]

However, Cohen was explicit that his ES guidelines serve as a very rough rule of thumb to be used as a last resort, and only apply where there is no theoretical or practical rationale available for determining the importance of an ES magnitude. In most cases the importance of a magnitude of an ES should be defined in relation to the qualities of a particular variable and the context of the study. Many textbooks do recognise Cohen's actual position and provide a nuanced discussion of the relationship between ES magnitude and importance.

> "Cohen's diffidence toward criteria for characterizing effect sizes was, in part, a consequence of his opinion that the size of an effect depends on what is being studied." (Dattalo, 2008, p. 41.)[7]

In summary, although many authors do recommend Cohen's rough and ready rule of thumb regarding the interpretation of the size of ESs with little qualification, there is an increasing recognition within the literature that the relationship between the magnitude of an ES and the substantive importance of an ES is by no means straightforward. However, it should be emphasized that such a recognition does not take into account the more fundamental issue that the ES magnitude is itself influenced by study design.

Sample size and ES

Many textbooks devote considerable space to the influence of sample size on sampling distributions of basic statistics such as the mean. Sample size is also frequently discussed in relation to ESs. However, the most frequent assertion regarding ES and sample size is that

[6] See also 10, 20, 23, 26, 32, 38, 68, 77, 90, 97, 112, 115, 116, 119, 130, 135.
[7] See also 3, 19, 22, 51, 81, 83, 88, 104, 111.

ESs are *not* influenced by the size of samples; indeed, many textbooks go further and state that one of the great advantages of ESs is that, unlike *p* values, ESs are uninfluenced by sample size.

> "Effect size is a standardized value that indicates the size of a difference with respect to a measure of spread, but is not affected by sample size." (Nolan & Heinzen, 2007, p. 543.)[8]

This view is misleading with regard to the actual relationship between ESs and sample size. Whether or not a particular test statistic value (e.g. a *t* value) is statistically significant is of course entirely dependent on sample size. For a given difference in means, with the same standard deviations, the *t* value as well as the *p* value change as a function of the increase in sample size. It is quite true that the same does not apply for the ES. For a given difference in means, with the same standard deviations, the actual value of the ES statistic does *not* change as sample size changes.

However, sample size does profoundly affect ES (Fan, 2001) because, although it does not change the magnitude of the effect, it does influence the reliability and precision of the estimate. Furthermore, small samples tend to produce more extreme ES estimates, which, when combined with bias for reporting significant results, can result in spurious ES estimates being introduced into the literature. A small number of textbooks do acknowledge that small samples can lead to unreliable estimates of ES.

> "That is an effect size may be less reliable with smaller numbers – say under 25." (Orlich, Harder, Callahan, Trevisan & Brown, 2012, p. 95.) [9]

[8] See also 43, 57, 84, 85, 87, 92, 100, 120, 129.

[9] See also 9 and 91

Furthermore, the size of the sample influences both the estimate of the means and standard deviations due to the central limit theorem. Confidence intervals of ESs do directly reflect the influence of sample size (i.e. everything else being equal, the confidence interval of the ES will be smaller the larger the sample size). Some general textbooks contain information about confidence intervals for ESs (e.g. Beins, 2012; Little, 2013), but extensive discussions of confidence intervals are largely restricted to books about the new statistics (e.g. Cumming, 2012), books specifically about ESs (e.g. Ellis, 2010) and books on meta-analysis (e.g. Grissom & Kim, 2014). It should also be noted that confidence intervals of more straightforward statistics, such as the mean, are widely misinterpreted (Hoekstra, Morey, Rouder & Wagenmakers, 2014). Sample size has another direct effect on ES. There is a considerable literature that suggests that ESs are in many cases inflated when small sample sizes are used (Baguley, 2009), particularly with reference to some very common measures of ES such as Cohen's *d* (Nagakawa & Cuthill, 2007). Other common measures of ES such as eta squared also tend to be an over estimate of variance explained with small sample sizes (Levine & Hullet, 2002). There is also evidence that large sample sizes may lead to underestimates of ES (Bakan, 1966). This is simply not discussed in general textbooks (Rutherford, 2012, is an exception). In summary, the message from most textbooks is that, despite much evidence to the contrary, sample size is not an important factor in the interpretation of ESs.

Categorisation of ESs: Unstandardized and standardized ESs

There are many ways to categorize ESs, but the most basic categories are standardized ESs that scale differences and relationships according to the distributions (such as *d* and *r*) and unstandardized ESs which are in essence the descriptive statistics (such as the raw difference between means). The majority of textbooks use the term ES to refer to standardized ESs and define ESs with reference to standardized ESs.

> "Effect size is expressed in standard deviations and plotted on a normal distribution curve." (Jarvis, 2006, p. 193.) [10]

Many textbooks do not even mention unstandardized ESs. Those that do include are Baguley, (2012), Brace, Kemp and Snelgar, (2016), Robins, Fraley and Krueger, (2009) and Wilcox (2010); however, the vast majority of textbooks only describe standardized ESs, and actually define ESs with reference to standardized ES.

When different types of ES are mentioned, the reference is usually to variance accounted for ESs versus difference measures using the standard deviation.

> "Effect Size: A statistically significant outcome does not give information about the strength or size of the outcome […….]. Statisticians have proposed many effect size measures that fall mainly into two types or families, the *r* family and the *d* family." (Morgan, Leech, Gloeckner & Barrett, 2013, p. 101.)

The neglect of unstandardized ES has been noted in a very small number of textbooks.

> "Unfortunately, the current emphasis on standardized effect sizes has led some researchers to omit the original metric means in published articles. Both the standardized and unstandardized results should be reported even though a choice is made to focus on one or the other in the text describing the experimental outcome." (Huitema, 2011, p. 14.)

This neglect of unstandardized ESs is widespread despite the fact that the APA task force report on the reporting of statistics recommends the use of unstandardized ESs where

[10] See also 37, 54, 59, 64, 75, 95, 96, 99, 103, 118, 122, 131.

possible (Wilkinson & APA Task Force on Statistical Inference, 1999). In review papers examining the reporting of ESs (e.g. Sun, Pan & Wang, 2010) there is also a tacit assumption that ES only refers to standardized ES, because almost all published studies include raw ESs, for example means of groups. One important problem associated with standardized ESs is that wildly different data sets can produce precisely the same standardized ES value. For example, a *t*-test with a very small absolute difference between the means and small standard deviations could produce precisely the same Cohen's *d* value as a *t*-test with a very large difference between the means and large standard deviations. The absolute magnitude of the difference between the means and the distributions around those means could be crucial to the psychological interpretation of the result. Baguley (2009) provides an excellent summary of the principal advantages of unstandardized ESs in many contexts; the most striking advantage is that unstandardized ESs are entirely free from the influence of the experimental control of the size of the variance (Baguley also notes that the units are meaningful and the calculations are straightforward). Given the importance and advantages of unstandardized ESs in many circumstances, their neglect in textbooks is potentially very damaging for the discipline.

What are the appropriate indices of ES?

Most authors neglect the issue of appropriateness of different measures of ES for different research designs. The vast majority of texts only mention *d* or *r*. There are of course exceptions; for example Morgan, Leech, Gloeckner and Barrett (2013) provide an excellent and lengthy account of the relevant issues. One reason for the neglect may be the fact that there is no consensus on this issue (Levin & Robinson, 1999). The major decisions regard whether to select differences measured in terms of standards deviation (standardized mean differences, the *d* family) or indices expressing the ES as a percentage of variance (proportion of the variance explained, the *r* family). Correlational studies almost always use some variant

of the *r* family because the ES is also the test statistic, whereas simple tests of difference are more likely to use the *d* family. There is no consensus on which is the preferred option, and from a conceptual point of view there is little difference. However, there is almost a complete neglect of the issue of biased and unbiased indices of ESs. There are of course exceptions where there is recognition of the importance of the selection of the effect size.

> "In addition, the choice of effect size index (e.g. Cohen's *d* or Hedge's *g*) is a critical decision…" (Hulsizer & Woolf, 2009, p. 141.)[11]

Authorities recognize that many of the standard measures of ES are biased and suggest alternatives (Baguley, 2009). For example, Hedge's *g* is a more unbiased estimate than *d*, whereas epsilon squared and omega squared are regarded as more unbiased estimates of population ESs than the more commonly used eta squared. There are also some very coherent arguments for matching particular designs to particular measures of ESs (e.g. Olejnik & Algina, 2003). All of these issues are widely neglected in textbooks. There is also no agreement about how to calculate the ES for even very basic statistical analysis. For example, there is no consensus about the correct formula to use for calculating the ES for a repeated measures *t*-test (Cumming, 2012). The arguments surrounding the latter (or any repeated measure design, see Baguley [2012] and Bakeman [2005] for guidance) are by no means straightforward. They focus on the best estimate of variability; for example, one method simply uses the standard deviation of the variable measured at time one, whereas another alternative uses the standard deviation of the differences between the pairs of scores.

Remaining definitions of ES

[11] See also 63 and 138

We have discussed a range of issues covered in textbooks regarding ES which have included references to how ESs are defined. For example, many textbook authors have defined ESs in terms of their practical significance. The discussion of standardized and unstandardized ESs also involves defining the nature of ESs. So here we add some definitions not covered elsewhere.

A number of journal articles have already made the point that there are different and often contradictory ways in which ESs are defined. Nakagawa and Cuthill (2007) point out that ES can refer to: (a) "a statistic which estimates the magnitude of an effect" (e.g. *r*); (b) "the actual values calculated from certain effect statistics" (e.g. *r* = .3); or (c) "a relevant interpretation of an estimated magnitude of an effect from the effect statistics" (e.g. "medium") (p. 593). Several authors define ES with reference to the null hypothesis.

> "…Whereas a test of statistical significance is traditionally used to provide evidence (attained *p* value) that a null hypothesis is wrong, an effect size (ES) measures the degree to which such a null hypothesis is wrong (if it is wrong)." (Grissom & Kim, 2005, p. 5.) [12]

Defining ES in relation to falsification is placing ES firmly within the framework of NHST. Placing ESs within this context could be seen as contradictory as many of the definitions used in previous sections (see *Sample Size and ES* in particular) define ESs in contrast to hypothesis testing and *p* values. Such a view may be best summarised by this quote from Greer and Mulhern (2002):

[12] See also 11, 14, 42, 45, 47, 60, 67, 78, 107, 125, 133, 134, 136, 137.

> "…effect size – the strength of the systematic relationship in the data that is independent of statistical significance." (p. 264.)[13]

Other textbooks give no real explanation of ESs but merely provide a minimalist description in the manner in which the calculation of the mean might be described.

> "A standardized way to describe the difference divides it by the estimated standard deviation for each group. This is called the effect size." (Agresti & Finlay, 2008, p. 200.)[14]

These explanations are not really helpful to the novice or experienced researcher as they provide no explanation or context. Other authors mention the term ES in passing but offer no real introduction or explanation for the term (e.g. Greasley, 2007).

The final definition is ES as an objective measure of the observed effect.

> "An effect size is simply an objective and (usually) standardized measure of the magnitude of observed effect". (Field, 2018, p. 113). [15],[16]

The claim for the objective nature of ES is particularly interesting. The *p* value of .05 has been regarded as an objective criterion for making decisions about whether or not a result is of importance and the authors are suggesting the substitution of ESs for *p* values for making this decision. However, as discussed previously, there is increasing recognition that the relationship between the magnitude of a particular ES and its substantive meaning is

[13] See also 21 and 44
[14] See also 2, 5, 7, 8, 18, 31, 35, 36, 39, 41, 46, 52, 55, 59, 69, 105, 108, 117.
[15] In fairness to the authors – they then go on to provide a more nuanced account and discuss the importance of interpretation of ESs in the context of particular experiments.
[16] See also 33 and 34.

problematic; what constitutes a small, medium or large ES is entirely dependent on context. Therefore, the claim of objectivity of the ES is particularly problematic.

Quantitative summary of coverage of ESs in textbooks

In summary, our review of textbooks has revealed that there is a widespread neglect of ESs. Furthermore, the most common accounts to be found regarding ES are frequently misleading. Good coverage of ESs is the exception rather than the rule. In figure 1 we summarise our review of the textbooks. We note whether six domains of interest have been covered accurately, inaccurately or minimally/not at all in the textbooks we have reviewed. The domains refer to the major problems covered in the preceding text: (1) the extent to which ES is presented as a simple index of practical real world effects without regard to the influences of experimental design (this domain is labelled 'practical' in the figure); (2) ES is interpreted as being associated with particular values of small, medium and large regardless of design and context (labelled 'magnitude' in the figure); (3) sample size is presented as having little or no effect on ES (labelled 'sample size' in the figure); (4) there is little or no coverage of unstandardized ESs (labelled 'unstandardized' in the figure); (5) the effect of selection of ES measure is not discussed (labelled as 'measure' in the figure); and (6) ES is defined in terms of reference to the null hypothesis (labelled as 'null' in the figure).

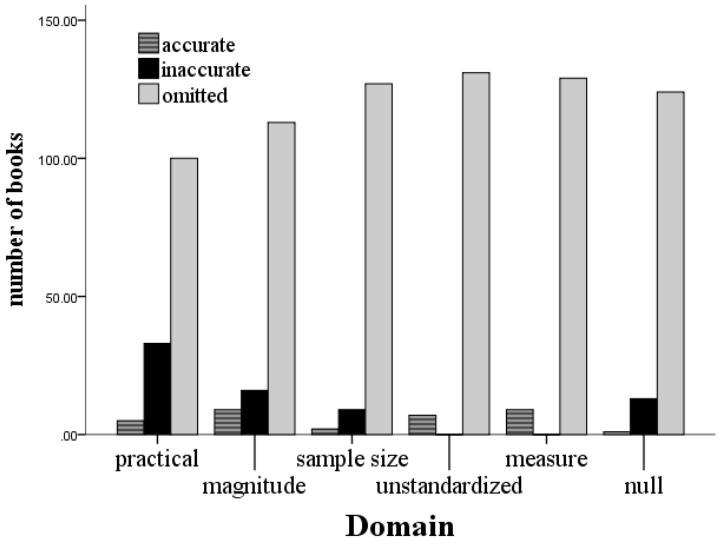

Figure 1. Number of books with accurate, inaccurate or no/minimal coverage of six domains regarding ESs (n = 138 textbooks)

The most striking aspect of figure 1 is that most textbooks devote very little coverage to ESs at all. However, the most common description to be found is that an ES is a measure of real world practical importance.

Conclusions

We have presented evidence that many textbooks make strong claims that ESs provide unproblematic information about the practical, real world importance of experimental results; in fact, many textbooks claim that this is the primary purpose of ESs. Furthermore, we present existing arguments that ESs in experimental settings is profoundly influenced by several aspects of experimental design entirely under the control of the experimenter. The crucial connection that is not routinely made is that if the ES is under the control of the experimenter, it cannot be a straightforward measure of real world, practical significance. Furthermore, we also suggest that standardized ESs in experimental settings are never an unbiased estimate of the real work effect (whereas causality may generalise from the experiment to the "real world"), as, by definition, experiments control extraneous variables and hence reduce the error term, and therefore increase the magnitude of the ES. We have the

paradox that the more artfully controlled the experiment (and many experimentalists would instinctively associate high levels of control as the hallmark of quality), the more inaccurate the estimate of the real world effect. The review of textbooks has highlighted a number of misunderstandings regarding ESs that are crucial for researchers when interpreting their results. We have shown that with regard to ESs there is widespread neglect of the impact of sample size, unstandardized vs. standardized measures, the relationship between ES magnitude and substantive meaning, and selection of appropriate ES measures. We would suggest that researchers (and indeed textbook writers) pay particular attention to the following issues.

1) Researchers should appreciate the advantages of unstandardized ESs in many contexts.
2) Rules of thumb equating particular values of small, medium and large ESs should be treated with caution and ESs should be always interpreted with regard to the local context of the study.
3) There is recognition that experimental design is a key determinant of the magnitude of ES.
4) There should be a full discussion of selection of the appropriate ES measure with particular regard to sample size and the design of the experiment.
5) ESs should be considered in the context of experimental and correlational studies.
6) There should be extensive discussion about how a variety of sources including descriptive statistics, graphs, *p* values, ES and confidence intervals should all be used in an active interpretation of the results.

We think it worth speculating on the reasons for the nature of the narratives about ESs presented in most textbooks. One of the great appeals of NHST is the – spurious – sense of objectivity it provides about the criterion for judging whether or not a result is legitimate for

interpretation. The same tendency would seem to be in operation with reference to ESs. This can be seen in the desire to define what is a small, medium and large ES regardless of context. Indeed, as we have already noted, at least one textbook makes the explicit claim that ESs provide an "objective" (Field, 2018) index of the importance of a result. If there is no external, objective metric that can be used to justify that a result is important, the alternative is that scientists will have to make an argument for the importance of their results on a study by study basis, using a wide variety of metrics from the results. This has, perhaps, the appearance of a more subjective and less rigorous science. Although the paradox is that there is strong evidence that "harder" sciences are much less reliant on statistical significance than "softer" sciences, and the harder sciences rely much more on the active interpretation of graphs and other descriptive information, using them as rhetorical devices to support arguments (Arsenault, Smith & Beauchamp, 2006; Smith, Best, Stubbs, Johnston, & Archibald 2000; Smith, Best, Stubbs, Archibald & Roberson-Nay, 2002).

Finally, textbook versions of a discipline can be a powerful block to change. Compelling critiques of NHST have been around for many decades, but NHST is still the sine qua non of almost all statistics textbooks in the behavioural sciences. Textbook versions of science take on a life of their own and seem remarkably resistant to change. However, textbooks could also be a powerful agent of change given their widespread influence. We suggest that there needs to be a concerted effort to improve the coverage of ESs in textbooks to make sure that researchers do not simply substitute one single index of importance, the significance level, for another, the ES (Pek & Flora, 2018). The American Statistical Association makes the excellent point that "[n]o single index should substitute for scientific reasoning" (Wasserstein & Lazar, 2016, p. 12).

References

Abbot, M.L. (2010). *Understanding educational statistics using Microsoft Excel and SPSS*. Somerset, US: Wiley.

Agresti, A. & Finlay, B. (2008). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River: Pearson International.

Anthony, D. (2011). *Statistics for health, life and social sciences*. Cophenagen: Book Boon. https://doi.org/10.7748/cnp.10.10.6.s3

Arsenault, D. J., Smith, L. D., & Beauchamp, E. A. (2006). Visual inscriptions in the scientific hierarchy: Mapping the "treasures of science". *Science Communication, 27(3),* 376-428. https://doi.org/10.1177/1075547005285030

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100(3),* 603-617. https://doi.org/10.1348/000712608x377117

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave Macmillan.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66(6),* 423-437. https://doi.org/10.1037/h0020412

Bakeman, R. (1992). *Understanding social science statistics: A spreadsheet approach*. New Jersey: Lawrence Erlbaum Associates, Inc.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37 (3),* 379-384. DOI: 10.3758/BF03192707}

Bausell, R. B., & Li, Y. F. (2002). *Power analysis for experimental research: a practical guide for the biological, medical and social sciences.* Cambridge: Cambridge University Press.

Beins, B. C. (2012). *APA style simplified*. Somerset US: Wiley.

Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science, 18(1)*, 1-32. http://dx.doi.org/10.1214/ss/1056397485

Berkman, E. T., & Reise, S. P. (2011). *A conceptual guide to statistics using SPSS*. Thousand Oaks: Sage. https://doi.org/10.4135/9781506335254

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of *d* for range restriction and unreliability. *Organizational Research Methods, 4(1),* 46-61. https://doi.org/10.1177/109442810141003

Boring, E.G. (1919). Mathematical versus scientific significance. *Psychological Bulletin, 16(10)*, 335-338.

Brace, N., Kemp, R., & Snelgar, R. (2016). *SPSS for Psychologists (and everybody else)* (6th ed.) Basingstoke: Palgrave.

Card, N. A. (2015). *Applied meta-analysis for social science research.* New York: Guilford Publications. https://doi.org/10.1111/insr.12011_17

Chandler, R.E. (1957). The statistical concepts of confidence and significance. *Psychological Bulletin, 54(5),* 429-430. https://doi.org/10.1037/h0041052

Cohen, B. H. (2013). *Explaining psychological statistics*. Somerset, US: Wiley.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65(3),* 145-153.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49(12),* 997-1003.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education*. London: Routledge.

Coolican, H. (2009). *Research methods and statistics in psychology*. Brighton: Psychology Press.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis.* New York: Russell Sage Foundation.

Cortina, J.M., & Landis, R.S. (2009). When small effect sizes tell a big story, and when large effect sizes don't. In C.E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences* (pp. 287-308). London: Routledge.

Costa, R. E., & Shimp, C. P. (2011). Methods courses and texts in psychology: "textbook science" and "tourist brochures". *Journal of Theoretical and Philosophical Psychology, 31(1),* 25.

Costall, A., & Morris, P. (2015). The "textbook Gibson": The assimilation of dissidence. *History of Psychology, 18(1)*, 1-14. https://doi.org/10.1037/a0038398

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25(1),* 7-29.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., & Wilson, S. (2007). Statistical reform in psychology is anything changing? *Psychological Science, 18(3),* 230-232. https://doi.org/10.1037/e538102007-001

Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64(3),* 138-146.

Cunningham, C. J., Weathington, B. L., & Pittenger, D. J. (2013). *Understanding and conducting research in the health sciences.* Chichester: John Wiley & Sons.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology.* Harlow: Pearson Education.

Dattalo, P. (2008). *Sample-size determination in quantitative social work research*. Oxford: Oxford University Press.

Davis, A. S., & D'Amato, R. C. (Eds.). (2010). *Handbook of pediatric neuropsychology*. New York: Springer Publishing Company.

Davis, C. (2013). *SPSS step by step: essentials for social and political science.* Bristol: Policy Press.

Dearholt, S., & Dang, D. (2012). *Johns Hopkins nursing evidence-based practice: Models and guidelines*. Indianapolis: Sigma Theta Tau.

Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science, 4(1),* 21. https://doi.org/10.1111/j.0956-7976.2005.01538.x

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(2),* 170-177.

Dunn, D.S. (2015) *The Oxford handbook of undergraduate psychology education*. Oxford: Oxford University Press.

Efron, S.E., & Ravid, R. (2013). *Action research in education. A practical guide*. New York: Guilford Press.

Elliott, A. C., & Woodward, W. A. (2014). *IBM SPSS by example: A practical guide to statistical data analysis.* Thousand Oaks: SAGE.

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

Evans, J. (2007). *Your psychology project: The essential guide*. Thousand Oaks: Sage Publications.

Evans, A. N., & Rooney, B. J. (2013). *Methods in psychological research*. Thousand Oaks: Sage Publications.

Faherty, V. E. (2007). *Compassionate statistics: Applied quantitative analysis for social services (With exercises and instructions in SPSS).* Thousand Oaks: Sage Publications.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94(5)*, 275- 282.

Fern, F. E., & Monroe, K. L. B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23(2)*, 89 - 105.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. London: Sage.

Field, A., & Hole, G. (2002). *How to design and report experiments*. London: Sage.

Field, A., Miles M. J. & Field, Z. (2012). *Discovering statistics using R* (5th ed.) London: Sage.

Fink, A. (2012). *Evidence-based public health practice*. Thousand Oaks: Sage.

Fitzpatrick, J.J., & Kazer, M.W. (2011). *Encyclopedia of Nursing Research* (3rd ed.). New York: Springer Publishing.

Forshaw, M. (2007). *Easy statistics in psychology: a BPS guide*. Chichester: Wiley-Blackwell.

Frick, R.W. (1994). Defending the statistical status quo. *Theory and Psychology, 9*, 183-189.

Gamst, G., Meyers, L. S., & Guarino, A. J. (2008). *Analysis of variance designs: A conceptual and computational approach with SPSS and SAS.* Cambridge: Cambridge University Press.

George, D., & Mallery, P. (2016). *IBM SPSS Statistics 23 step by step: A simple guide and reference.* London: Routledge.

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

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. What you always wanted to know about significance testing but were afraid to ask. In D. W. Kaplan (Eds.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391-408). Thousand Oaks: Sage.

Gochman, D. S. (Ed.). (2013). *Handbook of health behavior research II: Provider determinants.* New York: Springer Science & Business Media.

Godshall, M. (2015). *Fast facts for evidence-based practice in nursing: Implementing EBP in a nutshell* (2nd ed.). New York: Springer Publishing Company.

Gorard, S. (Ed.). (2007). *Combining methods in educational and social research*. New York: McGraw-Hill Education.

Gravetter, F. J., & Wallnau, L. B. (2014). *Statistics for the behavioral sciences* (8th ed.) Andover: Cengage Learning.

Greasley, P. (2007). *Quantitative data analysis using SPSS*. New York: McGraw-Hill Education.

Greer, B., & Mulhern, G. (2002). *Making sense of data and statistics in psychology*. Basingstoke: Palgrave.

Grimm, L. G. (1993). *Statistical applications for the behavioral sciences*. Chichester: Wiley.

Grimm, L. G., & Yarnold, P. R. (1995). *Reading and understanding multivariate statistics*. Washington: American Psychological Association.

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research. Univariate and multivariate applications.* London: Routledge.

Hair, J.F. Jnr., Black, W.C., Babin, B.J., Anderson, R.E., & Tatham, R.L. (2006). *Multivariate data analysis* (6th ed.). New Jersey: Pearson.

Halai, A, & Clarkson, P. (2015). *Teaching and learning mathematics in multilingual classrooms.* New York: Springer.

Hanneman, R. A., Kposowa, A. J., & Riddle, M. D. (2012). *Basic statistics for social research*. Chichester: John Wiley & Sons.

Harris, P. (2008). *Designing and reporting experiments in psychology*. New York: McGraw-Hill Education (UK).

Hartas, D. (2015). *Educational research and inquiry: Qualitative and quantitative approaches.* London: Bloomsbury Publishing.

Hayes, N., & Stratton, P. (2013). *A student's dictionary of psychology*. London: Routledge.

Hinton, P. R. (2014). *Statistics explained* (3rd ed.) New York: Routledge.

Hoekstra, R,, Morey, R.D., Rouder, J.N. & Wagenmakers, E.J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin Review, 21*, 1157-1164. DOI 10.3758/s13423-013-05723-3

Holt, N. & Walker, I. (2009). *Research with people: Theory, plans and practicals.* Basingstoke: Palgrave Macmillan.

Howell, D.C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.) Belmont: Thompson Wadworth.

Howitt, D., & Cramer, D. (2014). *Introduction to statistics in psychology*. Harlow: Pearson Education.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *The Journal of Experimental Education, 61(4),* 317-333. https://doi.org/10.1080/00220973.1993.10806593

Huck, S.W. (2013). *Reading statistics and research* (6th ed.) New York: Pearson.

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Chichester: John Wiley & Sons.

Hulsizer, M. R., & Woolf, L. M. (2009). *A guide to teaching statistics: Innovations and best practices.* Chichester: John Wiley & Sons.

IBM Corp. Released (2017). *IBM SPSS Statistics for Windows, Version 25.0*. Armonk, NY: IBM Corp.

Jarvis, M. (Ed.) (2006). *Sport psychology: A student's handbook*. Boca Raton: Taylor and Francis.

Johnson, B., & Christensen, L. (2008). *Educational research: Quantitative, qualitative, and mixed approaches.* Thousand Oaks: Sage.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice-Hall.

Kirch, W. (Ed.). (2008). *Encyclopedia of Public Health: Volume 1: A-H*. New York: Springer Science & Business Media.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation*. Thousand Oaks: Sage.

Lee, S., Dfinis, M. C. D. S. N., Lowe, L., & Anders, K. (2016). *Statistics for international social work and other behavioral sciences.* Oxford: Oxford University Press.

Lenth, R.V. (2007). Statistical power calculations. *Journal of Animal Sciences, 85,* E24-29. doi:10.2527/jas.2006-449.

Lenz, E. R., & Shortridge-Baggett, L. M. (2002). *Self-efficacy in nursing: Research and measurement perspectives.* New York: Springer Publishing Company.

Leong, F. T., & Austin, J. T. (2006). *The psychology research handbook: A guide for graduate students and research assistants.* Thousand Oaks: Sage.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review, 11(2)*, 143-155. https:// doi:10.1023/A:1022076425749

Levine, T. & Hullet, C. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28(4),* 612-625. DOI: 10.1111/j.1468-2958.2002.tb00828.x

Levine, N., Worboys, C., & Taylor, M. (1973). Psychology and the 'Psychology' Textbook: A social demographic study. *Human Relations, 26(4),* 467-478.

Levy, D. (1973). Psychological statistics: A teaching paradigm. *Bulletin of the British Psychological Society, 26(90),* 9 -12.

Little, T. D. (2013). *The Oxford handbook of quantitative methods, volume 1: Foundations.* Oxford: Oxford University Press.

Marschner, I. C. (2014). *Inference principles for biostatisticians*. Boca Raton: CRC Press.

Marsden, P. V., & Wright, J. D. (2010). *Handbook of survey research*. Basingstoke: Emerald Group Publishing.

Maruyama, G., & Ryan, C. S. (2014). *Research methods in social relations*. Chichester: John Wiley & Sons.

Marzano, R. J., Pickering, D., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement.* New York: Pearson.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison approach*. Belmont, CA: Wadsworth.

McGrath, R.E. (2011). *Quantitative models in psychology*. London: APA.

McNamara, J. F. (1997). *Surveys and experiments in education research*. New York: R & L Education.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46(4),* 806-834.

Melnyk, B., & Morrison-Beedy, D. (2012). *Intervention research: Designing, conducting, analysing and funding*. New York: Springer Publishing Company.

Mertens, D. M. (2014). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods.* Thousand Oaks: Sage publications.

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2013). *IBM SPSS for introductory statistics: Use and interpretation* (5th ed.). Abingdon: Routledge.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological methods, 7(1),* 105.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews, 82(4),* 591-605.

Newhouse, R. P., Dearholt, S. L., Poe, S. S., Pugh, L. C., & White, K. M. (2007). *Johns Hopkins nursing evidence-based practice model and guidelines*. Indianapolis: Sigma Theta Tau.

Nolan, S.A. & Heinzen, T.E. (2008). *Statistics for the behavioral sciences*. New York: Worth Publishers.

Norcross, J. C., Hogan, T. P., & Koocher, G. P. (2008). *Clinician's guide to evidence based practices: Mental health and the addictions*. Oxford: Oxford University Press.

Obiakor, F.E., Bakken, J.P., & Rotatori, A.F. (2010). *Current issues and trends in special education: Research, technology, and teacher preparation*. Bingley UK. Emerald Publishing Limited.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods, 8(4),* 434-447.

Oliveira, A., & Oliveira, A.G. (2013). *Biostatistics decoded*. Chichester: John Wiley & Sons.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92(3)*, 766-777.

Onwuegbuzie, A. J., & Levin, J.R. (2003). Without supporting statistical evidence, where would reported measures of substantive importance lead? To no good effect. *Journal of Modern Applied Statistical Methods, 2(1)*, 133-151.

Orlich, D. C., Harder, R. J., Callahan, R. C., Trevisan, M. S., & Brown, A. H. (2012). *Teaching strategies: A guide to effective instruction.* Andover: Cengage Learning.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical assessment, research and evaluation*, *8*(11), 1-7. http://edresearch.org/pare/getvn.asp?v=8&n=11

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation.* London: Dulau & Co.

Peat, J., & Barton, B. (2014). *Medical statistics: A guide to data analysis and critical appraisal* (2nd ed.). Chichester: John Wiley & Sons.

Pek, J., & Flora, D.B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods, 23 (2),* 208-225. http://dx.doi.org/10.1037/met0000126

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. *American Psychologist*, *34*(5), 373. https://doi.org/10.1037/0003-066x.34.5.373

Phakiti, A. (2015). *Experimental research methods in language learning*. London: Bloomsbury.

Polgar, S., & Thomas, S.A., (2013). *Introduction to research in the health sciences*. London: Churchill Livingstone.

Privitera, G.J. (2016). *Essential statics for the behaviorual sciences*. Thousand Oaks: Sage.

Ravid, R. (2010). *Practical Statistics for Educators*. Blue Ridge Summit, US: Rowman & Littlefield Publishers.

Richardson, J. T. E. (1996). Measures of effect size. *Behavioral Research Methods, Instruments, & Computers, 28(1),* 12–22.

Ringquist, E. (2013). *Meta-analysis for public management and policy*. Chichester: John Wiley & Sons.

Roberts, P. & Priest, H. (2010). *Healthcare research: A textbook for practioners and researchers.* Chichester: Wiley.

Robins, R. W., Fraley, R. C., & Krueger, R. F. (Eds.). (2009). *Handbook of research methods in personality psychology*. New York: Guilford Press.

Rosenthal, G., & Rosenthal, J. A. (2011). *Statistics and data interpretation for social work.* New York: Springer Publishing Company.

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

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2003). *Evaluation: A systematic approach.* Thousand Oaks: Sage publications.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57(5)*, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetico-deductive. In L.L. Harlow, S.A. Mulaik & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ: Erlbaum.

Rubin, A. (2012). *Statistics for evidence-based practice and evaluation*. Andover: Cengage Learning.

Rutherford, A. (2012). *ANOVA and ANCOVA*. Chichester: Wiley

Ruxton, G., & Colegrave, N. (2011). *Experimental design for the life sciences*. Oxford: Oxford University Press.

Sackett, P. R., Laczo, R. M., & Arvey, R. D. (2002). The effects of range restriction on estimates of criterion reliability: Implications for validation research. *Personnel Psychology, 55(4),* 807–825.

Sapp, M. (2006). *Basic psychological measurement, research designs, and statistics without math.* Springfield Illinois: Thomas.

Schumacker, R. E. (2014). *Learning statistics using R*. Thousand Oaks: SAGE.

Schwartz, B. M., Landrum, R. E., & Gurung, R. A. (2016). *An easy guide to APA style*. Thousand Oaks: Sage Publications.

Sloboda, Z., & Bukoski, W. J. (Eds.). (2003). *Handbook of drug abuse prevention*. New York: Springer.

Smith, L. D., Best, L. A., Stubbs, D. A., Johnston, J., & Archibald, A. B. (2000). Scientific graphs and the hierarchy of the sciences: A Latourian survey of inscription practices. *Social studies of science, 30(1)*, 73-94.

Smith, L. D., Best, L. A., Stubbs, D. A., Archibald, A. B., & Roberson-Nay, R. (2002). Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist, 57(10),* 749.

Smithson, M. J. (1999). *Statistics with confidence: an introduction for psychologists*. Thousand Oaks: Sage.

Soh, K. (2016). *Understanding Test and Exam Results Statistically: An Essential Guide for Teachers and School Leaders*. New York: Springer.

Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology, 15(1)*, 72-101.

Stangor, C. (2014). *Research methods for the behavioral sciences*. Belmont, CA: Wadsworth Publishing.

Steinberg, W. J. (2010). *Statistics alive!* Thousand Oaks: Sage Publications.

Stephens, R C., Scott, C.K., & Muck, R.D. (2002). *Clinical assessment and substance abuse treatment.* New York: State University of New York Press.

Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences* (3rd ed.) Abingdon: Routledge.

Stolerman, I. (Ed.). (2010). *Encyclopedia of psychopharmacology*. New York: Springer.

Stufflebeam, D. L., & Shinkfield, A. J. (2007). *Evaluation theory, models, and applications*. San Francisco: Jossey-Bass.

Suen, H.K. (1989). *Analyzing quantitative behavioral observation data*. New Jersey: Lawrence Erlbaum Associates.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology, 102(4),* 989.

Terrell, S. R. (2012). *Statistics translated: A step-by-step guide to analyzing and interpreting data.* New York: Guilford Press.

Thompson, B. (2008). *Foundations of behavioral statistics: An insight-based approach*. New York: Guilford Press.

Troidl, H., Spitzer, W.O., McPeek, B., Mulder, D.S., McKneally, M.F., Wechsler, A. & Balch, C. M. (2012). *Principles and practice of research: strategies for surgical investigators*. New York: Springer.

Urdan, T.C. (2005). *Statistics in plain English*. New Jersey: LEA.

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin, 72(3),* 204-213.

Walker, J., & Almond, P. (2010). *Interpreting statistical findings: a guide for health professionals and students*. London: McGraw-Hill Education.

Wasserstein, R.L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process and purpose. *The American Statistician, 70,* 129-133. http://dx.doi.org/10.1080/00031305.2016.1154108

Weinberg, S. L., & Abramowitz, S. K. (2002). *Data analysis for the behavioral sciences using SPSS.* Cambridge: Cambridge University Press.

Weiner, I. B., Schinka, J. A., & Velicer, W. F. (2012). *Research methods in psychology* (2nd ed.). Chichester: Wiley.

Welkowitz, J., Cohen, B. H., & Lea, R. B. (2010). *Introductory statistics for the behavioral sciences* (7th ed.). Chichester: John Wiley & Sons.

Whitney, P., & Ochsman, R.B. (1988). *Psychology and productivity*. New York: Springer.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. Practical Assessment, Research & Evaluation, 14(5), 1-9. ISSN-1531-7714

Wilcox, R. R. (2010). Fundamentals of modern statistical methods: Substantially improving power and accuracy. New York: Springer.

Wilcox, R. R. (2016). Understanding and applying basic statistical methods using R. Chichester: John Wiley & Sons.

Wilkinson, L. and Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and Explanations. American Psychologist, 54(8), 594-604. https://doi.org/10.1037//0003-066x.54.8.594