The reporting of effect size in Educational Psychology literature

Jennifer May Hampton

Supervised by: Catherine Fritz

April 2012

The reporting of effect size in Educational Psychology literature

ABSTRACT

This dissertation discusses issues of effect size in education, psychology and educational psychology literature. Reporting and interpretation of effect size estimates are discussed in terms of the purpose of individual research as conceptualised by Kirk (2001). Also discussed are issues surrounding the reporting and interpretation of null hypothesis significance testing (NHST); as are confidence intervals and matters of determining practical significance. The issues raised are also considered in terms of working within a research community, cumulative knowledge growth and reporting to a non-expert audience. The papers published in 2010 in the *Journal of Educational Psychology* and *Learning and Instruction* were surveyed to determine the reporting practices, specifically for those findings reported in the abstracts. The data reveal a large proportion of studies reporting, but not discussing, effect size estimates. A cumulative frequency was calculated from reported partial eta squared values, producing contextual guidelines for interpretation. These guidelines contrast with Cohen's (1988) but are similar to those found in other areas of psychology (Morris & Fritz, under review; 2012). Results are discussed in terms of trends in reporting and issues of interpretation. Overreliance on traditional methods as well as readily available effect size statistics calls for greater author engagement with the issue. Finally, comprehensive resources to guide researchers in these matters are presented.

KEY WORDS:	EFFECT SIZE	EDUCATIONAL PSYCHOLOGY	PRACTICAL SIGNIFICANCE	STATISTICAL SIGNIFICANCE	REPORTING PRACTICES

The reporting of effect size estimates in educational psychology

The reporting of effect size estimates has been advocated for quite some time (see Huberty, 2002). Relatively recent work, such as Cohen (1994), has inspired a renewed interest into the reporting, interpretation and usefulness of effect size estimates. This body of work has also raised issues of overreliance on traditional methods of analysis (Henson, Hull &Williams, 2010), such as null hypothesis significance testing (NHST). Policy has shifted to emphasise the importance of reporting, and to a lesser extent interpreting, effect size estimates. This is true both with regard to professional bodies (such as the APA, 2010) and journal editors (Grissom & Kim, 2012). There have been questions raised about the extent to which researchers have embraced, or at least reported, effect size estimates, as illustrated in a number of reviews (e.g., Fidler et al., 2005; Fritz, Morris & Richler, 2012; Osborne, 2008). This introduction examines why it is important that this translation to practice does happen, with reference to the purpose of individual research as well as wider implications and the philosophical considerations raised by a neglect to report effect size.

The Purpose of Research

Setting aside philosophical considerations of the purpose of research generally for the moment, let us consider purpose from an individual research level. Empirical, quantitative research may vary between the specific research questions being addressed. However, in the analysis and subsequent discussion of data, research is arguably trying to answer a few fundamental questions. These are usefully outlined by Kirk (2001) as:

1. Is the observed effect real?
2. If the observed effect is real, how big is it?
3. Is this effect big enough to be useful?

Briefly, quantitative researchers answer whether the observed effect is real by determining a sufficiently low probability of chance being responsible for the results obtained. This tends to be done through statistical significance testing, formally termed null-hypothesis significance testing. However, statistical significance testing does not indicate how large the observed effect is; in order to do this additional techniques need to be employed. These produce effect size estimates which can be used to describe how big an observed effect is. The final question asked concerns the usefulness of the observed effect: its practical significance. These three concepts – null hypothesis significance testing, effect size estimates and practical significance – are arguably the foundations of quantitative research analysis. Although seemingly straightforward, there are many issues surrounding their use, interpretations and limitations.

Null-hypothesis significance testing. Null-hypothesis significance testing (NHST), rejecting or retaining a hypothesis that is contrary to the experimental hypothesis based on a calculated p value, is possibly the most used statistical technique in psychology. Based on the characteristics of the sample data, the probability of obtaining those data if there was no effect is calculated. Researchers in psychology, education and most social sciences typically determine that the populations show the actual effect if that probability is less than .05. This practice of dichotomous testing for significance has been criticised, not least for the arbitrary nature of the .05 value (e.g., Rosnow and Rosenthal, 1989). Whilst there have been many others who are

disdainful of the practice of using NHST (not least Cohen, e.g. 1994), this dissertation is not an attack on the practice but examines the limitations and coinciding misinterpretations of the statistic.

NHST only tells part of the story (Kirk, 2001). Because it is testing for a rejection or retention of the null hypothesis, it does not address the truthfulness of the experimental hypothesis (Sun, Pan & Wang, 2010). A misconception that Cohen (1994) highlighted is that a 'successful' rejection of the null hypothesis equals an affirmation of the experimental hypothesis. In fact it does not even necessarily mean that the null hypothesis is false (Cohen, 1994). All a statistically significant result actually tells us is that the effect we have observed is not likely to be due to chance (Ellis, 2010). It is also worth being aware of the fact that often the null-hypothesis under consideration is the nil-hypothesis (i.e. that there is no relationship between, or no effect on, the variables being investigated). However as NHST is a function of sample size (Grissom & Kim, 2012), given a large enough sample, variables can be found to related to some extent no matter how tenuous the relationship. It follows that when the null is a nil-hypothesis it can always be rejected given a large enough sample (Cohen, 1994). This raises concern that such testing may be meaningless (e.g., Cumming, 2012; Grissom & Kim, 2012; Cohen, 1994).

Nickerson (2000) highlighted the misconception that a small p value is equal to a large effect. This can be observed in the literature with references being made to 'highly significant' p values. This misconception raises several issues, including the dichotomous nature of NHST, the size of the effect and consideration of the language used. A result is found to be statistically significant or non-significant depending on which side of the .05 alpha value the calculated p value falls. Leaving aside the arbitrary nature of this cut-off, the conclusion is always a dichotomous one; either the result is statistically significant or it is statistically non-significant. Therefore, referring to a p value as being 'highly' (or, conversely, 'nearly') significant is always inappropriate. Although this simple distinction may be adequate in some circumstances (Grissom & Kim, 2012), the often complex nature of issues being addressed by psychological research means that the simple dichotomy is inadequate (Sun et al., 2010) and the magnitude of the effect needs to be considered. A statistically significant p value may be interpreted as giving cause to believe the observed effect is 'real' but gives no clue as to the magnitude of the observed effect. An effect may be either statistically significant but trivial or statistically significant and important (Ellis, 2010), but the p value gives no indication of this (no matter how compelling the temptation when a small p value is observed). The very language used when referring to 'significance' leads to the danger of conflating practical significance with statistical significance, even by researchers themselves (Nickerson, 2000), not least because of the synonyms associated with the term. 'Significant' conjures up ideas of importance and noteworthiness, with 'non-significant' suggesting triviality. The danger of misinterpretation is apparent when these synonyms do not reflect the true nature of the data. In the very least, Thompson's (1996) suggestion that researchers must always refer to results as being 'statistically (non-) significant' rather than simply '(non-) significant' may help to avoid being misleading. If researchers themselves hold such misconceptions, and attendant misinterpretations, then perhaps one should question the quality of quantitative methodology courses. In addition, if these sorts of mistakes are being made by

expert researchers, it is likely that they are also being made by the wider non-expert audience.

Despite the limitations and dangers of misinterpretation of NHST, there is an overreliance on it (Henson et al., 2010) which has led some to argue that there is too strong a focus on significance tests alone when evaluating data and research. Testing for a rejection of a nil-hypothesis (as is often the case) does nothing to advance knowledge (Sun et al., 2010). This sentiment was also expressed by Kirk (2003) who warned of the danger of actually impeding scientific progress, with Schmidt (1996) warning of a retardation in the development of cumulative knowledge. Schmidt's argument focuses on the failure of traditional interpretations of research literature, based on statistical significance testing, to reveal the true meaning of data. This danger may be exacerbated by publication bias: journals do not tend to publish papers with statistically non-significant results (Thompson, 1996). This bias is also seen in authors, in that they are much less likely to even submit papers with statistically non-significant results (Grissom & Kim, 2012). Despite its opponents, NHST can be useful in the exploration of data. However, its importance is often exaggerated (Fan, 2001). Most importantly, it does not address the size of the effect, whether the results are noteworthy or not. Thompson (2007) argued that the real goal of research concerns the noteworthiness of results, which is related to the second question Kirk (2001) asked: How big is the effect?

Effect size estimates. Effect size estimates address some of the limitations of null-hypothesis significance testing but also have some limitations of their own, discussed below. Throughout the discussion, reference is to effect size estimates; they describe the effect in the sample and provide a point estimate for the size of the effect in the population (Fritz et al., 2012). Primarily effect size estimates answer the magnitude question as they describe the size of the observed effect; the extent to which the observed results differ from the null-hypothesis (Vacha-Haase & Thompson, 2004). In addition to aiding the individual piece of research, not least in their "critical" role in informing interpretation of the observed effect (Henson, 2006; p.604), effect size estimates are beneficial to the wider research community. Not only does reporting effect size estimates help later researchers in their power analyses, such estimates also play a crucial role in meta-analysis. Linked to the criticism levelled at NHST in retarding the development of cumulative knowledge (Schmidt, 1996), effect size estimates can enhance such knowledge accumulation. When not given in the research and where data given allow it, meta-analysts calculate effect size estimates for those studies included in their analyses. However, research is excluded from such meta-analyses where effect size estimates are not given and the data reported are insufficient. Clearly, calculating effect size estimates in primary analysis leads to more accurate meta-analyses (as asserted in Grissom & Kim, 2012). Staying with meta-analysis for a moment, Thompson (2008a) elaborated on Cohen's (1994) assertion that it is a misconception that the complement of NHST is replication of a significant p value. Thompson argued that findings can only be found to be replicated if the effect size found is consistent across studies. This can be said to be the case if the effect size can be generalised when variations in design and measurement are taken into account. This has links with a form of meta-analysis that searches for homogeneity across studies (Grissom & Kim, 2012), a method that is criticised by some (for example, Hunter and Schmidt, 2004).

Efforts have been made to increase awareness of the benefits of calculating and reporting effect sizes, through articles such as those by Cohen (e.g. 1994) and Thompson (e.g. 2008a), and through publication policies of journals (see Thompson's online list of journals that require effect size reporting) as well as professional bodies (e.g. the APA, 2001, 2010). Furthermore, as indicated in the introduction to the 5th Publication Manual Edition, the APA (2001; p. 5) state that failure to report effect size estimates can be considered to be a 'defect' of design and/or reporting of research. Failure to report is also detrimental to the accumulation of knowledge for the discipline, when consideration is given to the implications for limiting comprehensive meta-analyses. Despite these reasons for reporting effect size there are still a surprising number of quantitative papers that do not do so; a more detailed discussion of trends in reporting and policy will be given in a later section. Thompson (1999) suggested one reason why initial attempts to make effect size estimates reporting a standard practice may not have worked. He criticised the APA's (4th edition, 1994) manual 'encouraging' authors to include effect size estimates, suggesting that an encouragement in a manual full of requirements infers less importance regarding this practice. Whilst this may be true, and the APA has taken steps to address this in subsequent manuals, there are several other reasons why authors might not address the issue of effect size.

There are many effect size estimates to choose from, the most simple of which is the difference between means, a method advocated by Baguley (2009). Whilst this method describes the data, it does not describe the relationship and is not without problems (Fritz et al., 2012). Setting this simple metric aside, there are literally dozens of methods to choose from (Sun et al., 2010) which in itself means researchers may be overwhelmed. In addition, whilst there are broad categories that the various methods fit into, not all of them do so easily (Grissom & Kim, 2012). Both of these issues illustrate that simply deciding on which effect size estimate to use in the first place is not a simple task and can be confusing. This complexity also highlights the importance of adequate training (Schmidt, 1996) which may be lacking, as suggested by the variety of misconceptions already highlighted regarding NHST. Indeed textbooks have traditionally paid little (if any) attention to effect size estimates (Thompson, 2007). If researchers are unfamiliar with effect size estimates, it should not be surprising if they not comfortable using them (Fritz et al., 2012). This unfamiliarity may arise through a lack of training but also through a lack of exposure to others using them in the literature. Even with training, if a method is not often used or seen being used, it may not be fluently understood (Thompson, 2002). This lack of understanding, in turn, may lead to a lack of use and a cycle of failure to report may persist.

Of course it may not be through a lack of understanding that effect size estimates are not being reported. It may be that authors are reluctant to report effect size estimates for a variety of reasons, perhaps because they think that NHST, simply establishing an effect, is sufficient. The editors questioned by Cumming et al. (2007) noted "authors' resistance" (p. 231) to reform despite editors' overt encouragement. This resistance may have reflected authors' underlying thinking about the use of effect size in relation to NHST, a relationship that will be explored in greater detail in a later section. Interestingly, Cumming et al. found a contrasting position from the authors themselves, who expressed a support of effect sizes. They displayed a

willingness to contemplate change in reporting practices but cited editorial policy as a barrier.

Some researchers are critical of the proponents of effect size estimate reporting for overlooking the limitations inherent with the practice. Critics, such as Onwuegbuzie, Levin and Leech (2003), have been unhappy with the lack of discussion of these limitations, criticising the exclusive focus on benefits, often accompanied by a simultaneous attack on NHST. Included in the criticisms levelled by Onwuegbuzie et al. (2003) is the sheer amount of choice of effect size estimates to use. Rather than focus on the confusing nature of this issue, they raised the interesting (yet cynical) possibility that with such choice it is possible to choose a self-serving measure that reflects the authors' own agenda. Another consideration to bear in mind is that although effect size estimates are generally independent of sample size (unlike NHST), in some cases the effect size bias can be influenced by the size of the sample. Grissom and Kim (2012) were keen to point out that this bias arises rarely and has only a slight bearing on outcomes. As with any other examination of data, the estimate can also be influenced by the soundness of the underlying design and features of the particular analysis. Indeed, effect size estimates depend on means and standard deviations of the sample under investigation and so some effect size estimates may vary across studies (Sun et al., 2010), although these variations can be overcome by meta-analysis (Henson, 2006). Given a small sample size, effect size estimates tend to over-estimate the effect in the population (Grissom & Kim, 2012). This problem may be exacerbated by a publication bias towards the publication of statistically significant results. Statistical significance is a function of sample size and effect size, and so the likelihood is that small studies that are statistically significant also have large effect sizes. Reporting effect size estimates for all results, significant and non-significant, may in theory help avoid this over-inflation of population estimates (Grissom & Kim, 2012). However without addressing the publication bias for papers with statistically significant results, over-inflation of population effects across research literature will persist to some extent.

The final consideration is the limitation of what effect size estimates actually tell us. Whilst giving a description of the size of the effect observed (Fritz et al., 2012), effect size estimates do not show how useful the effect actually is. As Henson (2006) succinctly stated, effect size estimates in themselves are not inherently meaningful. Consideration needs to be given to the wider audience particularly. Onwuegbuzie et al. (2003) observed that effect size estimates may be particularly meaningless to consumers. Whilst this lack of meaning to the wider audience can arguably be overcome by proper interpretation on the part of the author, Onwuegbuzie et al. suggested that other measures entirely may be more appropriate; these alternative measures are discussed in the next section.

Despite these limitations, effect size estimates do give researchers a basis from which the usefulness or practical significance of the effect may be inferred (Ellis, 2010). Researchers need to consider what the values that they have obtained actually mean in order for inferences to be made. Indeed without interpretation, reporting effect size estimates obtained "adds little to a report of research" (Grissom & Kim, 2012; p. 5).

Issues of reporting. Kirk's (2001) questions for individual research imply a sequential approach. If (1) a statistically significant result is found then (2) the size of the effect should be determined. This implies that if a non-significant result is found then there is no need to investigate issues of effect size. There are some proponents of this approach. Wilkinson (1999) advised that a p value should always be accompanied by an effect size estimate. However it is common practice not to give a p value when the result is non-significant and so this advice could be taken to support sequential reporting. Onwuegbuzie et al. (2003) argued that because both NHST and effect size estimates have their limitations they should be used together, in a sequential nature. They argue that this combination reduces the danger they perceived of effect size estimates misrepresentation; addresses misinterpretation of p values; and increases consistency between analysis and conclusions. Whilst including an effect size estimate may address a misinterpretation of a small p value meaning a large effect (Nickerson, 2000), it is debateable whether effect size estimates are misrepresented or whether a practice of reporting in this way would increase consistency. Despite these last two points, the idea that these methods complement one another was supported by Fan (2001), who also warned that they address separate questions and so do not substitute for one another. This line of reasoning suggests that addressing effect size is not dependent on finding a statistically significant result.

Researchers such as Thompson (2007) have argued that effect sizes need reporting for all results, both statistically significant and non-significant. This has some connection with the fact that NHST is a function of sample size, as well as considerations of meta-analysis and the accumulation of knowledge. Sun et al. (2010) argued that both NHST and effect size estimates need to be reported because of the discrepancies that can occur between the two measures. An effect may be found to be statistically non-significant but large or may be statistically significant but small. These discrepancies have different implications (as discussed in Fritz et al., 2012). As statistical significance is a function of sample size, a large but statistically non-significant result simply suggests that the study needed more power, something that might have been calculated before the results were obtained. Less commonly, a statistically significant result may have a small effect size which suggests that caution should be used with such results. Although this is an uncommon occurrence, except in the realm of large scale surveys, Fritz et al. (2012) give the somewhat startling example of a correlation of .1 having (one-tailed) statistical significance with a sample size of just 272. Reporting both NHST and effect size estimates highlights such discrepancies. Unfortunately however, as Alhija and Levy (2009) pointed out, such discrepancies are "frequently not reported, not interpreted, and mostly not discussed or resolved" (p. 245).

As well as bringing to light any discrepancies that may occur, and implications these discrepancies suggest, reporting both NHST and effect size estimates is useful for meta-analyses. Reporting effect size estimates generally improves the accuracy of such analyses. Including smaller effect size estimates that may arise from non-significant results also avoids over-inflation of effect size estimates over multiple studies (Grissom & Kim, 2012). This brief overview shows that some researchers are loyal to the dominance of NHST in the analysis of data. Addressing both, regardless of which results are obtained, has benefits for the individual piece of research, clarity for the audience, and can lead to more accurate meta-analyses.

This approach seems to be the most supported, at least in theory if not in practice (Cumming et al., 2007). Finally there are those who argue that NHST should be abandoned altogether. Schmidt (1996) is one of these proponents who argue that NHST should be replaced with effect size estimates and confidence intervals. Although this dissertation will not consider these issues in depth, confidence intervals are discussed briefly below.

Confidence intervals. What Kirk's (2001) questions fail to address is how confident the researcher can be as to the accuracy of their estimates. A limitation of both NHST and effect size estimates is that they are both point estimates and give no indication as to accuracy of the estimate. Using confidence intervals allows inferences to be made as to the precision of the estimates calculated (Ellis, 2010). In essence they do the same job as NHST but also give the range in which the effect is likely to lie. Kirk stated that it is "hard to understand why researchers have been so reluctant to embrace" them (2001; p. 214). However, it may be for similar reasons that effect sizes are not reported. There may be a similar barrier of low confidence and unfamiliarity stopping researchers from reporting them. Alternatively, or simultaneously, inadequate teaching or textbook guidance might explain avoidance of confidence intervals. Confidence intervals can also be calculated and used to determine the range and accuracy of effect size estimates. Despite the APA (2001) stating that providing confidence intervals for the effect size estimates is the best reporting strategy, it is not common practice. Again this may be for similar reasons regarding the reluctance to report effect size estimates themselves. Fritz et al. (2012) also observed that confidence intervals for effect size estimates are not intuitive in that they are not symmetrical around the point estimate. These considerations are worth bearing in mind but because of the infrequency of their use confidence intervals are not discussed further.

Practical significance. The final question Kirk (2001) asked of individual research is whether the observed effect is big enough to be useful. The usefulness of the effect can also be referred to as the substantive (Onwuegbuzie et al., 2003) or practical significance. In this case 'significance' really does reflect its synonyms; it is concerned with the importance and implications of the effect found. This question addresses what Thompson (2007) argued is the real goal of research, whether results are noteworthy. As discussed previously, it is important to remember that whilst practical significance can be informed by effect size, they are not synonymous with one another (Grissom & Kim, 2012). Interpretation of effect size estimates obtained is needed to establish practical significance (Sun et al., 2010). This is particularly important for the wider audience of the research. All research reporting needs to be clear, concise and understandable and particular consideration should be given to those who are to apply the implications of such research. Grissom and Kim (2012) highlighted the importance of making effect size estimates accessible with their warning that not doing so "is a kind of withholding of information" (p. 9).

The question then arises as to how to interpret effect size estimates. The guidelines set out by Cohen (1988) are the most commonly used, and misused, interpretation method employed in quantitative psychological research. There are limitations with using any kind of fixed benchmark, such as these, not least because of their arbitrary nature (as stated by Harris, 2008, amongst many others). Cohen (1988) himself warned that his guidelines were "proposed… with much diffidence, qualification and

invitations not to employ" and that they were "no more reliable than my own intuition" (p. 532). Although seemingly simplifying a complex issue, these guidelines are misused because researchers tend to not heed Cohen's warnings and follow them 'blindly' (Thompson, 2008a). Applying such guideline benchmarks, and accompanying labels of 'small', 'medium' or 'large', does not actually address the practical significance of the effect. Simply labelling effects in such a way, without further elaboration of what they actually mean, implies value that may not be appropriate. A 'small' effect may be interpreted by the reader as having little importance, but without considering the context of the study, the importance of the outcomes, or the previous research context this cannot be known. All three of these issues need consideration by the researcher when determining the practical significance of the results (Henson, 2006). It is possible to marry these two approaches to some extent. Morris and Fritz (under review; 2012) present guidelines that are calculated from effect size estimates reported in previous research from specific research domains. This allows the 'small', 'medium' and 'large' descriptors to be calculated and applied in a non-arbitrary way.

Leaving aside fixed benchmarks, there are a variety of methods used across the literature (Thompson, 2008a) that can be employed to help determine the practical significance of the effect they have observed. Many of these use external references such as clinical and even economic indicators, fulfilling Onwuegbuzie et al.'s (2003) hope for researchers to use methods that go "beyond 'internally referenced' effect sizes" (p. 39). With specific reference to educational research, Harris (2008) gave an outline of an economics approach based on cost-effectiveness, taking into consideration financial and other restraints. It is an approach commonly used in health research but not one that seems to have been taken up by the educational research community as yet. Narrowing the focus yet further, Hill et al. (2008) considered educational intervention particularly. They raised the point that there are no universal guidelines, so researchers must fight the temptation to use simple benchmarks. Referring to 'substantive significance', they argue that using normative growth expectations or policy-relevant gaps in achievement between certain groups (e.g. socio-economic, ethnic, even gender) is a more salient way of conceptualising the implications of observed effects. They also reiterate the need for, and the usefulness of, considering previous effect size estimates in the research.

Practical significance is not an easy thing to determine. It takes thought and deliberation over a wide range of considerations. It is a vital part of research is to determine how useful the effect observed is not least because of the responsibility researchers have to their audience (Grissom & Kim, 2012) and those who the results of research may impact upon. Whilst there are some useful methods of assessing the usefulness of findings, caution should be taken when using any fixed benchmark (e.g., those proposed by Cohen, 1988). Finally, with all generalisations and implications taken from research, reporting researchers must not go beyond the limits of the design of the study (Olejnik & Algina, 2003).

Trends and Policy

Whilst there is a long history of literature concerning effect size estimates (Huberty, 2000), the focus of this dissertation is on discussions stemming from Cohen's work. The emphasis on effect size reporting has been rapidly increasing over recent years (Grissom & Kim, 2012). This is reflected in the changing policies of journals and

bodies, such as the American Psychological Association (APA) and the American Educational Research Association (AERA) as well as their British counterparts: the British Psychological Society (BPS) and the British Educational Research Council (BERA). The APA responded to increasing concern raised by researchers by instructing a 'task force on statistical inference' whose recommendations were published (Wilkinson, 1999). On the basis of these recommendations (and possibly also in response to Thompson's (1999) criticism of the wording of the 4th edition guidelines), the 5th publication manual (2001) superseded the encouragement of reporting an effect size estimate to a necessity; it is "almost always necessary to include some index of effect size reporting" (p. 25). The BPS reiterates this with their manuscript requirement that "in normal circumstances, effect sizes should be reported" (item 4j). More recently, the 6th edition of the APA (2010) publication manual reaffirms the importance of reporting "effect sizes, confidence intervals, and extensive descriptions" because they "are needed to convey the most complete meaning of the results" (p. 33). However, this guidance is slightly different to making the results comprehensible in terms of Grissom and Kim's (2012) definition. Grissom and Kim assert the importance of results being accessible to the wider audience, through interpretation. This assertion is reiterated in the AERA (2006) guidelines. Their guideline that "qualitative interpretation of the index of the effect that describes its *meaningfulness*" (p. 37, italics added) should be included, represent perhaps the most comprehensive policy in terms of answering all of Kirk's (2001) questions. All this policy translates into submission guidelines for journals published by the above bodies, although in practice APA style is the standard practice for many if not most psychology journals. There has been a general increase in journal editors recommending, if not requiring, effect size estimate reporting (Grissom and Kim, 2012). Some editors have clearly taken the recommendations to heart. For example, Murphy (1997) writing as editor of the *Journal of Applied Psychology* stated there had to be good reason not to include an effect size estimate (and at the time of writing no one had come up with one yet). He stated the benefit of allowing readers to assess the true meaning of results but did not mention interpretation on the part of the researcher as necessary. In terms of journals that require effect size estimate reporting, Thompson (2008b) lists 24 journals that state this practice as a necessity. However, despite this requirement, there are those who question whether these requirements are enforced (e.g., Fritz et al., 2012; Grissom and Kim, 2012).

Studies investigating effect size reporting practices are found across various specialisms within the discipline of psychology. Investigating trends within clinical psychology between 1993 and 2001, Fidler et al. (2005) found that effect size reporting only increased a little and that this increase was only found in studies employing ANOVA analytical methods. It should be noted, however, that the period under investigation was before the APA introduced their revised guidance on the matter. A later study by Fritz et al. (2012) into practices of papers published in the *Journal of Experimental Psychology: General* found that fewer than half of the papers reported an effect size measure. Educational psychology has shown similar reporting practices, despite Harris' (2008) assertion that this is now a common reporting practice and an important advancement in educational research. Several studies bear contrary evidence to this assertion of the apparent widespread nature of such reporting. Writing in the same year, Osborne (2008) and Matthews et al. (2008) published results from their surveys of trends in educational psychology (1969-1999) and gifted education (1996-2005) respectively. Both found very limited

increases in reporting practices, with Osborne concluding that practices had effectively "failed to change" (as asserted in the title of the study). Reviews of both educational research (Keselman et al., 1998) and teacher education (Zientek, Capraro and Capraro, 2008) a decade later found few instances of effect size estimate and confidence interval reporting. Finally, Sun et al. (2010), reviewing education and psychology, found similar results with effect size estimates infrequently reported and even less frequently interpreted. However, there were suggestions that at least some researchers were engaging with the concept, with some instances of discrepancies between significance testing and effect size estimate results.

Philosophical Considerations

To really understand why the issues raised here are important it is necessary to frame them in the wider context of the purpose and ethos underlying psychological research. Whilst there is a range of different specialisms within the discipline, Bray (2010) argued that there is a common identity shared by psychologists. This identity is defined by the core of methods and scientific rigour that is shared. Although there has been a resurgence of 'common sense' and naïve realism in recent years (Lillenfield, 2010), it is important to remember that scientific thinking is not intuitive. This is particularly true for issues explored in this dissertation, as Kahneman (2011) asserted "even good statisticians were not intuitive statisticians" (p. 5). A specific example of the somewhat counter-intuitive nature of such issues is that confidence intervals are not symmetrical around effect size estimates (Fritz et al., 2012) because their distributions are not symmetrical, normal distributions. Another problem with intuitive statistical reasoning is the law of small numbers presented by Amos and Kahneman (1971): researchers are inclined to think that a randomised sample drawn from a population will be highly representative, and they are correct if the sample is sufficiently large. But smaller samples have greater variability than larger samples (Slavin & Smith, 2009) and they are more likely to produce extreme results. Matters such as these need serious thought, consideration and understanding. It follows that concerns over the quality of training of under- and post-graduates, as raised by Henson et al. (2010) and demonstrated in the misconceptions regarding NHST (e.g. Nickerson, 2000), must be given due attention.

Whilst scientific thinking is not intuitive, neither must it be wholly objective. Indeed, despite common conceptions of science as entirely objective, Thompson (1996) argued that "science is inescapably a subjective business" (p.28). Researchers make judgements constantly when conducting research; for example when creating research questions, evaluating existing research, designing studies, and deciding which analytical methods to employ. However, there is a commonly held belief in the 'objective nature' of scientific enquiry. Perhaps the distinction between bias and judgement is not sufficiently clear, and so one might assume that to be subjective or use subjective judgement is to be unavoidably biased. Clearly this is not the case; researchers use their expert, subjective judgement in a variety of unbiased decision making situations. For example, for any given study there may be several possible unbiased designs to choose from. The decision a researcher makes regarding which design to use is fundamentally an unbiased yet subjective one. The nature of NHST, of knowing exactly what the p value obtained means and the ability to make an easy, definitive conclusion based upon it, is objective and therefore might be seen

to be more ‘scientific’ (Kirk, 2001) than other, less dichotomous approaches. However, Berger and Berry (1988) argued that even within statistics “objectivity is not generally possible” (p. 165). Making decisions about interpreting the effect size estimates obtained and the practical significance of results requires “subjective judgement” (Grissom & Kim, 2012, p. 4). This subjective element may not sit well with what some researchers regard as the ‘objective’ nature of scientific enquiry, which may go some way to explaining some apparent reluctance to interpretation in the literature (e.g. Fritz et al. 2012; Sun et al., 2010).

Science is “the business of discovering replicable effects” (Thompson, 1996; p. 28). As such, research needs to be reported in such a way that enables it to be replicated, allowing the possibility of findings to be duplicated (Stanovich, 2011). It is this duplication of results that allows findings to become publicly verified (Stonovich, 2011). It is crucial therefore to consider what is meant by replication and duplication. Considering the limitations of NHST, it is arguably inappropriate to consider replication of a statistically significant result as duplication of findings. For example, if two studies investigating the same issue both find a statistically significant effect, but with widely different effect sizes, it is questionable whether these represent a replication of findings. Duplication of the size of the effect, however, clearly demonstrates replication (Thompson, 2008a). It is this replication, along with extensions, of research which allows cumulative knowledge to grow (Stanovich, 2011). Reporting effect size estimates is beneficial, in terms of cumulative knowledge growth, as it enables meta-analysts to more accurately (and efficiently) conduct their work (Grissom & Kim, 2012). These kinds of reporting practices also benefit the research community in that they help other researchers with their power analyses. This notion that psychologists are working within a community, towards understanding that can only really be obtained on a larger scale than individual research, highlights some of the reasons for good reporting practices. As well as being replicable, findings must be reported in such a way that enables extension and criticism (Stanovich, 2011). The process of peer review is one mechanism that allows research to be subject to such criticism. The peer review process also demonstrates the importance given to working within a research community, as it is peers that review work, and public verifiability, as papers are scrutinised before publication. Research can also be refused if it is found to be ‘trivial’ (Stanovich, 2011). If an effect is found to be statistically significant but very small, this may suggest triviality or at the very least caution (Fritz et al., 2012). However, if the effect size estimates are not being reported, discrepancies in effect size estimates will be less apparent (as found in Sun et al., 2010). Not reporting effect size estimates also makes it unlikely that this sort of failure to replicate will be picked up by peer reviewers.

Psychologists have a responsibility to the audience of their research. This includes the wider, ‘non-expert’ audience as well as other researchers. When reporting findings it is expected that this is in a clear, concise manner that conveys their “complete meaning” APA (2010; p. 33). It is generally accepted (in policy at least) that in order to do this, some measure of effect size needs to be incorporated. However, in order for effect size estimates to be understandable, some interpretation must be done. It has been well argued that fixed benchmarks (such as Cohen’s) are inappropriate. Indeed they may well be meaningless to a non-expert audience. Interpretation therefore needs to be conducted in terms of the wider context, both

previous research and wider implications of the findings. Surely it is the experts – the researchers themselves – are best placed to conduct such interpretation, which will be beneficial to audience interpretation of the piece. The language used must also be borne in mind when reporting findings. In terms of the non-expert audience, the danger of misinterpretation of findings, due to the synonyms associated with significance, is high. This is especially worrying when considering the finding by Nickerson (2000) that conflation of statistical and practical significance is committed by researchers themselves. The issues of language and interpretation become particularly relevant when considering those applied research areas that are likely to impact on and be implemented by non-experts. Encouragingly, applied psychology journals are better than average when it comes to reporting effect size estimates (Morris & Fritz, 2011).

Present Research

The present research focuses on reporting practices in educational psychology. This follows other research investigating practices within both applied psychology (e.g., Fritz et al., 2012; Morris & Fritz, 2011), education research (e.g., Keselman et al., 1998; Zientek et al., 2008), and educational psychology (e.g., Osborne, 2008). By its very nature, the implications of any applied research must be considered. Indeed, published research can have direct implications on teaching practice, as illustrated by the impact cognitive psychology has had on the teaching of reading (Stanovich, 2001). There is also a need to consider less direct implications. An example provided by Thompson (2007) demonstrates the need to consider the nature of educational psychology interventions. An affected change in learning may prove incremental over time, thus impacting participants' future, as well as present, learning. Going beyond Thompson, this indirect impact has the potential to help shape the citizens (and leaders) of the future and so is potentially of concern to the whole of society. By its very nature education is a societal concern. As such, both issues of the usefulness of research and making research understandable are particularly salient. Consideration must also be given to the fact that most educators work with finite resources (in terms of time, people and finances), often within publicly funded institutions. In education, the expected benefit of a given intervention, for example, must be considered against the investment needed to implement it. This can only be an informed judgement if effect size estimates are reported, and is likely to be better informed when interpreted by the researchers themselves.

The present research surveys current educational psychology literature, investigating the extent to which the considerations discussed here are represented. The literature is examined for each of the elements of Kirk's (2001) three questions for individual research: NHST, effect size estimates and practical significance. Of additional interest is whether any discrepancies between NHST and effect size estimates are reported and/or addressed. Reporting of power analyses will suggest that the issue of effect size is at least being given some attention. Finally, although not the main focus of this study, confidence intervals are counted in the survey for they represent an uncommon practice that may indicate engagement with a questioning of the overreliance on traditional techniques such as NHST (e.g. Henson et al., 2010).

Method

To investigate the reporting of effect size in recent educational psychology literature, two journals were chosen from this field. Educational psychology journal impact factors for 2010 were obtained through the JCR social sciences portal. The journal with the highest impact factor score was *Child Development*. This journal was not chosen for use in this study because of the nature of the content. The journal addresses a wide range of areas pertaining to child development, without a particular focus on education. Of the remaining top few journals, the *Journal of Educational Psychology* (second highest score) and *Learning and Instruction* (third highest score) were chosen. An overview examination of the content revealed that both had published an adequate proportion of studies using quantitative analytical methods. The two journals also offered a good comparison with one another. The *Journal of Educational Psychology* is an APA paper, and instructions to submitting authors required the reporting of effect size. *Learning and Instruction*, on the other hand, does not state any requirement for effect size reporting.

All papers published in 2010 from the *Journal of Educational Psychology* (volume 102) and *Learning and Instruction* (volume 20) were obtained. This amounted to 61 papers from the *Journal of Educational Psychology* and 48 papers from *Learning and Instruction*. All papers were downloaded. Papers that were a review, commentary or editorial were excluded from analysis. Of the remaining papers a note was made of the area of psychology being investigated, the topic covered, and the 'major' finding(s). Major findings were identified as those the authors chose to report in the abstract. By choosing to report them in the abstract, it is assumed that it is these findings that are the main contribution of the paper in question. These can be considered the primary outcomes that Wilkinson (1999) referred to and for which effect size estimates should be reported. A note was also made of the participants included in the research, and whether a power analysis had been conducted to determine the size of sample used.

The major findings for each paper were examined in detail. Examining the results sections of the papers, a note was made of the analysis used for each major finding. Those findings for which a qualitative or modelling approach was taken were not examined further. There are two distinct trends taken in analysing quantitative data, modelling approaches and ANOVA-type approaches (as illustrated in Henson, Hill and Williams, 2010). Although there is some literature on the calculation and reporting of effect size estimates for modelling approaches (e.g. Peugh, 2010), this study will focus on ANOVA-type approaches only. For the remaining findings a note was made of which statistics were used to describe the data and analysis. Specific attention was paid to the reporting of effect size statistics. When an effect size estimate was reported, a note was made of which statistic was reported, the specific value of this, and whether it was accompanied by confidence intervals. A note was made as to whether or not the effect size was discussed, and whether this discussion was in the results or discussion section of the paper. It was also noted whether the effect sizes were reported for all or some of the results, and whether these were statistically significant. Notes were gathered into a spreadsheet as the survey was taking place. Any other notes of interest were made as an aside to the main survey.

Once all the survey data from each relevant paper were collected, attention was paid to the frequencies of different analysis methods used, along with the reporting of effect size statistics and which effect size statistics were used. Aggregating the results for both journals separately enabled comparisons between the two. By recording the values of effect size statistics reported, it was also possible to calculate cumulative effect sizes for, in particular, partial η2. This process allowed Cohen's guideline values for 'small', 'medium', and 'large' effect sizes to be reconsidered in terms of the published sizes, enabling more meaningful interpretations of effect sizes found, as interpretations can be made in terms of a specific field, rather than Cohen's (1988) "intuition" (p. 532).

Results

The *Journal of Educational Psychology* published 61 papers in 2010 (volume 102). The majority of these papers used primarily quantitative statistical analyses, with just two reviews and one purely descriptive paper. Over half (36 papers) of the quantitative papers took a model-fitting approach to their analyses. A variety of methods were used, including confirmatory factor analysis, structural equation modelling and hierarchical linear modelling. These papers were excluded from further examination. Of the remaining 22 papers, inspection of the abstract for each paper revealed a total of 37 major findings. The most frequently used analysis used was multiple regression (used 15 times), followed by one-way AVONA (used eight times), and correlation (used six times). T-tests were used three times and multi-way ANOVA just twice. Proportions of the methods used in these remaining papers can be seen in Figure 1.

The majority of papers reported an effect size estimate of some kind, with only four findings not being supported by an effect size measure. I made a distinction between papers that merely reported an effect size estimate and those that also discussed or elaborated upon it. This distinction is particularly relevant for correlation and regression analyses as the coefficients reported are a measure of effect in themselves.

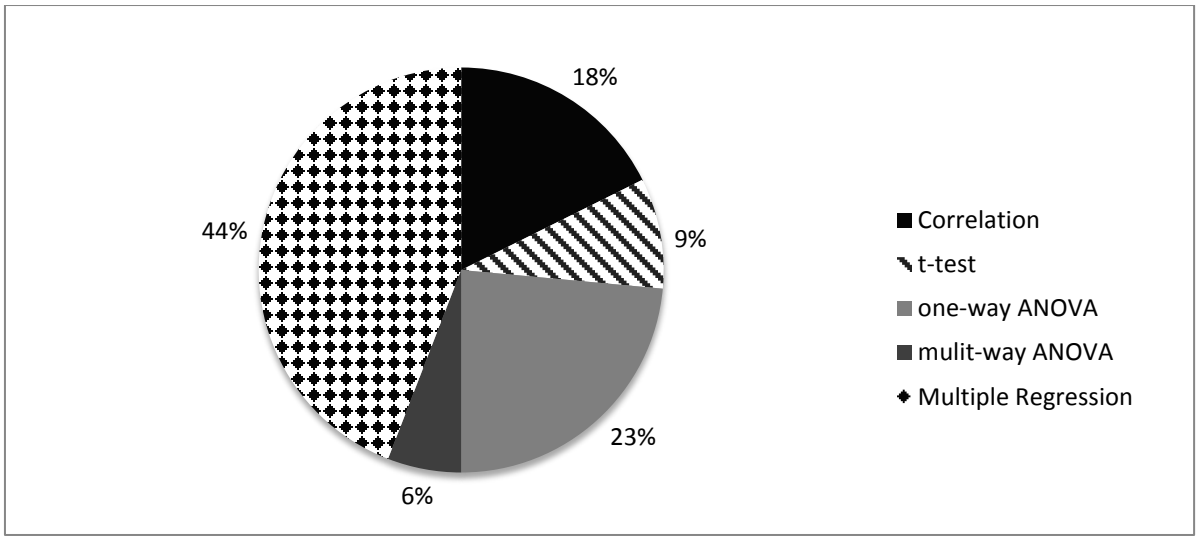

Figure 1. Analyses used in *Journal of Educational Psychology* papers.

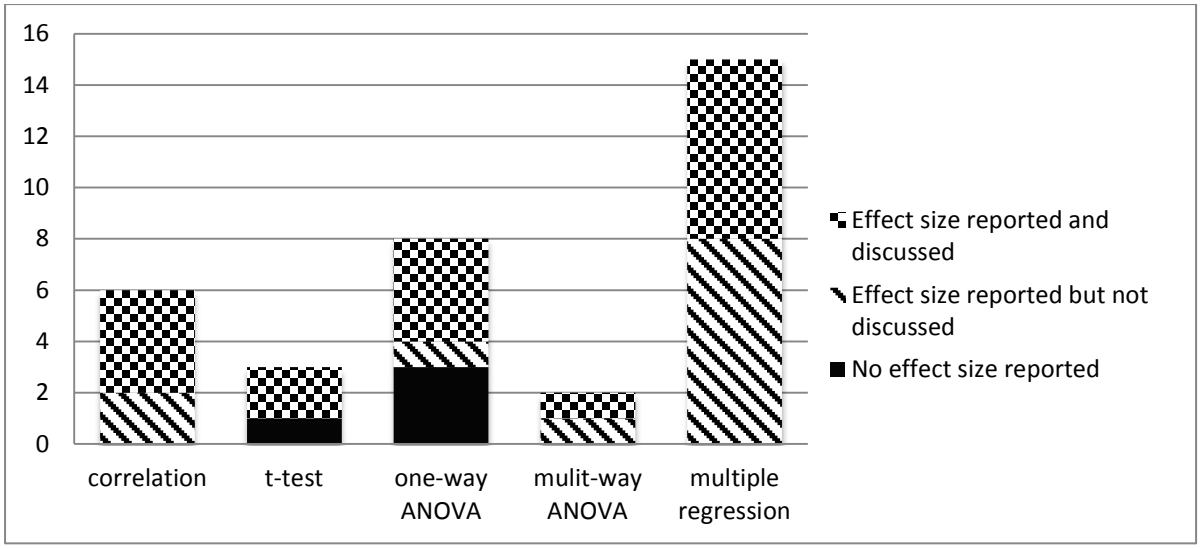

Figure 2. Reporting and discussion of effect size estimates in *Journal of Educational Psychology* papers.

Figure 2 shows that, generally, half of the reported effect size estimates were 'discussed'. In terms of effect size estimates reported for t-tests and ANOVA, discussion was usually simply in terms of attributing strength (small, medium or large) to the value obtained. These strength labels were given, in most cases, with direct reference to Cohen's benchmarks: with small = .01, medium = .06, large = .14. For effect size estimates reported in association with correlation and multiple regression analyses discussion was similarly brief. For these estimates, discussion centred on the amount of variance, or unique variance, in the dependent variable explained by the independent variable under investigation. The only effect size estimates that were discussed were those from statistically significant results.

Learning and Instruction published 48 papers in 2010 (volume 20). The majority of these papers also used primarily quantitative statistical analyses. However, seven papers took a model fitting approach, two papers took a qualitative approach, and seven papers were of a commentary or review nature. These sixteen papers were excluded from further analysis. Of the remaining 32 papers, inspection of the abstract and results sections of each paper revealed 53 major effects suitable for further inspection for the purposes of this study. Five other major effects were reported but employed either a qualitative or modelling approach, and in one case there were no results reported to support the claim in the abstract. Of the major effects under inspection, the most frequently employed method was multi-way ANOVA (25 times), followed by one-way ANOVA (15 times). T-tests were used four times and correlation just once. Proportions of each analysis used in terms of the papers under inspection are shown in Figure 3.

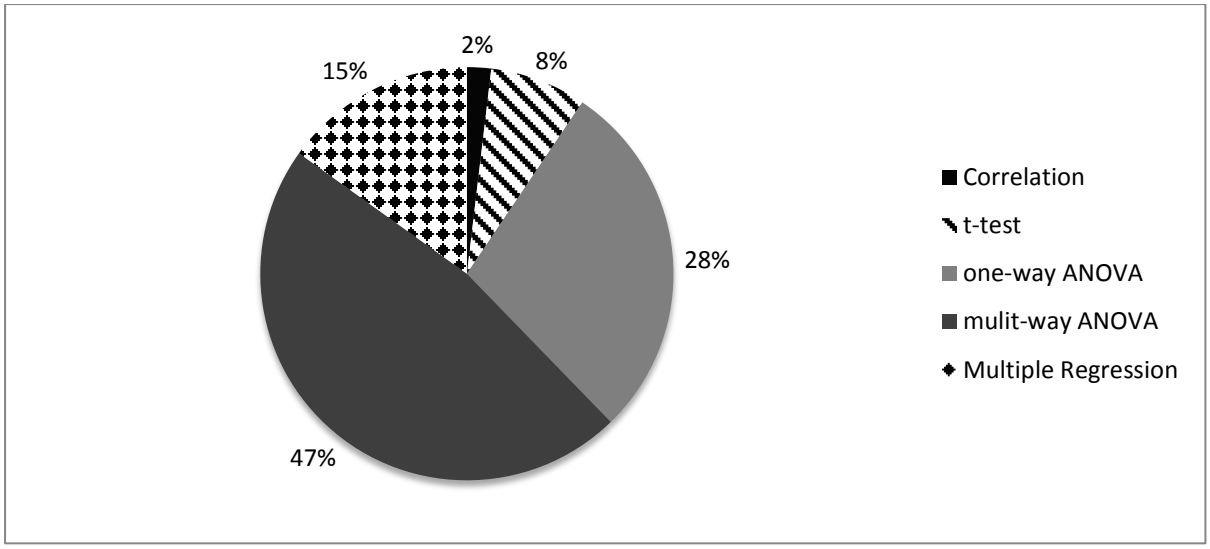

Figure 3. Analyses used in *Learning and Instruction* papers.

Effect size estimates were reported for all of the major findings examined in Learning and Instruction. Most effect size estimates reported were for statistically significant results. However, some effect size estimates were reported for statistically non-significant results in a few of the papers from this journal. This was generally when several values were being reported from the main analysis for the major finding. Although vigilant in reporting, authors did not often discuss the meaning of the effect size estimates, as shown in Figure 4. Similarly to effect size estimates reported in *Journal of Educational Psychology*, discussion was usually simply in terms of strength, with reference to Cohen's (1988) guidelines. Although less frequent for effect size estimates associated with multiple regression, there was some discussion in terms of proportion of unique variance explained in the dependent variable by the independent variable.

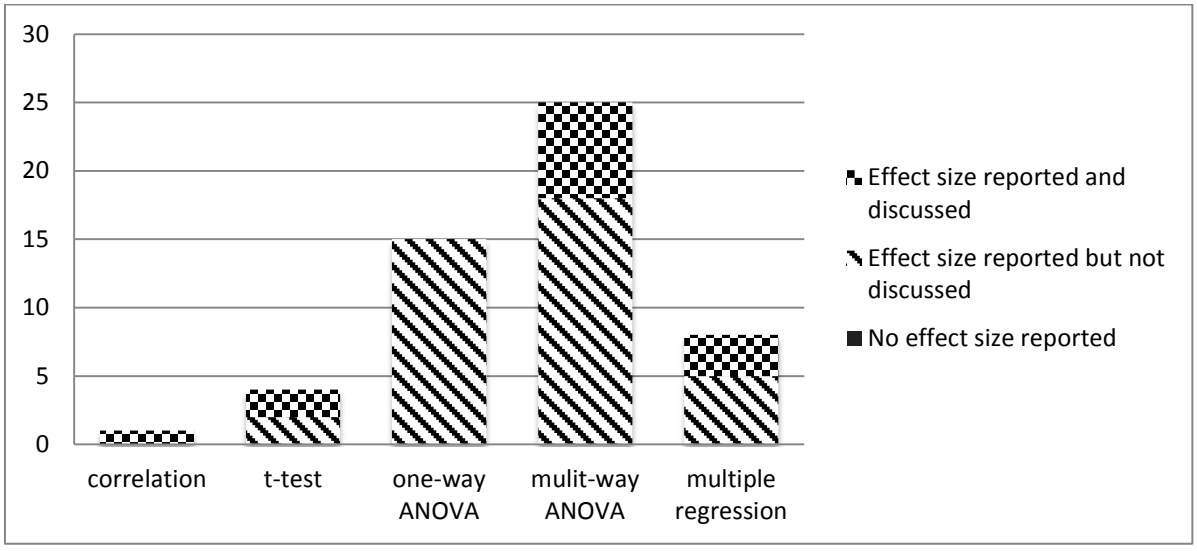

Figure 4. Reporting and discussion of effect size estimates in *Learning and Instruction* papers.

There were various types of effect size estimate reported between the two journals. Some types of estimate were used just once across the papers, including (but not limited to) Fisher's *z*, Cohen's f^2, adjusted R^2, and an instance of an odds ratio

method. The most frequently reported types of effect size estimate for ANOVA-type analyses were Cohen's *d*, η^2, and partial η^2. For correlation type analyses, *r*, *beta*, R^2, and R_{change}^2 were the most frequently reported types of estimate. Partial eta squared (partial η^2) was by far the most frequently reported types of effect size estimate overall, with 127 instances. When these reported effect size estimates were discussed it was generally brief and limited to noting strength in terms of Cohen's guidelines.

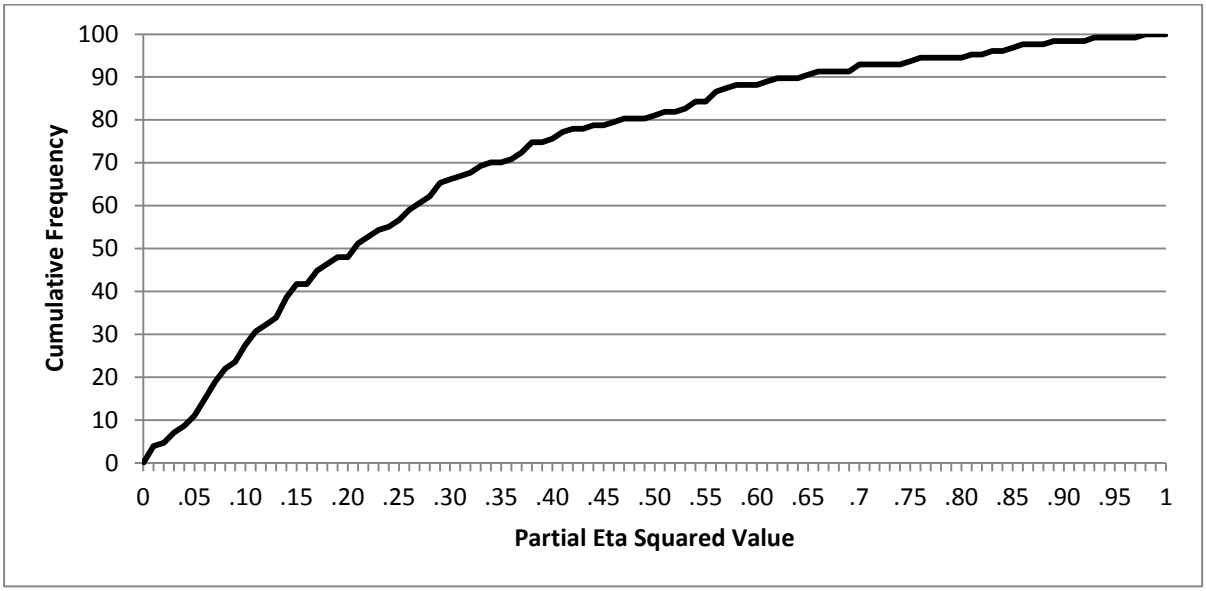

Figure 5. Cumulative effect size estimate frequency for partial eta squared reported in *Journal of Educational Psychology* and *Learning and Instruction*.

In order to offer a comparison to Cohen's guidelines, the partial eta squared values reported in surveyed papers were collated to produce a cumulative effect size reference, as shown in Figure 5. It is from this that new guidelines may be drawn. As these are calculated solely from reported effect size estimates in the specific field of educational psychology, they may be considered a non-arbitrary descriptor of effect size found in this area. As such, they also allow a measure against which the appropriateness of using Cohen's guidelines may be inferred. The first quartile may be thought of a small effect, the median as a medium effect, and the third quartile as a large effect. Table 2 shows the estimates from these data, in comparison with Cohen's estimates and those found by Morris and Fritz (under review; 2012) from a survey of applied cognitive psychology and memory literature from 2010.

Table 2
Estimated guidelines for determining strength of effect.

Effect Size	Educational Psychology	Cognitive and Applied	Memory	Cohen's Estimates
1st Quartile (Small)	.10	.08	.08	.01
Median (Medium)	.21	.18	.19	.06
3rd Quartile (Large)	.38	.42	.42	.14

Note: All except Cohen's are the quartiles calculated from published research.

No confidence intervals for effect size estimates nor power analyses were reported in any of the papers examined from either journal.

Discussion

In terms of reform of reporting practices, the results obtained are both encouraging and discouraging to greater and lesser extents. There are varying practices between the two journals in terms of statistical analyses preferred. The *Journal of Educational Psychology* demonstrated a tendency towards modelling and multiple regression, whilst *Learning and Instruction* tended towards using forms of ANOVA. These differing dominant analytical approaches may reflect the nature of the research conducted with the former tending to school-based, longitudinal studies and the latter taking a more traditional experimental approach. The two journals were seemingly similar in their reporting practices of effect size estimates however, with the majority of papers reporting some form of effect size measure. This contrasts with previous reviews that have found quite infrequent reporting of effect size estimates in related research areas (e.g., Matthews et al., 2008; Osborne, 2008). It was disappointing to note that the cases where no effect size estimates were reported appeared in the *Journal of Educational Psychology* in spite of it being an APA paper with standards that dictate that these kinds of estimates are reported. Indeed the journal itself instructs authors to include in the results "appropriate indexes of effect size or strength of relationship" (2003; p. 201) which overwhelmingly indicates that it is not editorial policy that is a barrier to reporting (as suggested by the authors surveyed in Cumming et al., 2007). Rather, lack of reporting in the face of such direct requirement suggests that some authors are choosing not to report effect size estimates for whatever reason be it unfamiliarity, neglect or an active measure for their own gains (perhaps in terms of hiding a small effect).

Encouragement that authors are engaging with the reporting of effect size might be taken from the results obtained in regards of reporting practices. Nevertheless, there are questions raised as to whether this reporting demonstrates real engagement. Certainly the majority of findings were supported by some measure of effect size. However, it may be that these estimates are reported simply because they are readily available. It must remembered that correlational statistics in themselves are a measure of strength of relationship and were counted as such in the research. Also, in terms of ANOVA, partial eta squared is provided with the output in SPSS and other statistical packages. The notion that effect size estimates are being reported without due thought and simply because of their ready availability is strengthened when considering other aspects of reporting observed here. Firstly, power analyses were not reported in any of the studies, suggesting that issues of effect size were not being considered by the authors. It is unlikely that researchers are calculating but not reporting such analyses, especially in light of the APA recommendation to "provide evidence that the study has sufficient power to detect effects of substantive interest" (2010; p. 30). Authors appear not be considering either issues of "sufficient power" or "substantive interest", as is discussed in more detail below. Secondly, no authors reported confidence intervals for any effect size estimates reported. Disappointingly, one paper that did actively engage in contemplation of effect size, using a method sought from a textbook addressing the issue (Schagen & Elliot, 2004), did not follow through with the instructions therein to

calculate (and report) confidence intervals associated with the effect size estimates. Thirdly, the way in which effect size estimates were discussed demonstrates a superficiality of engagement.

Fewer than half of the effect size estimates reported were discussed. Of those effect size estimates that were discussed, nearly all discussion was brief with reference to either variance explained (for correlational analyses) or Cohen's (1988) guidelines (for ANOVA-type analyses). This kind of limited discussion is often, understandably, disregarded by other reviewers and so may explain discrepancy between this research and other reviews; with this research finding a higher than usual proportion of studies engaging some discussion of effect size estimates. There was little evidence of authors trying to interpret or report these estimates further. Encouragingly, there was an instance of odds ratio being used. This is one technique to convert effect size estimates into more understandable form advocated by Ellis (2011). Despite this one example of a sign of engagement with interpretation issues, generally authors neglected to interpret effect size estimates reported.

The advantages of authors including effect size estimates, whether or not they are engaging with the issue fully, include the benefits to future research and the growth of cumulative knowledge. Reporting effect size estimates allows cumulative frequencies to be calculated, and thus provide a basis for new guidelines to determine effect strength, to be drawn (Morris & Fritz, under review; 2012). This approach was used to identify guidelines for educational psychology based on the studies used in this research; these empirically based guidelines are substantially larger than Cohen's (1988) intuitive guidelines. Indeed, for example a partial eta squared result of .12 would be considered large using Cohen's guidelines but small using these new, contextual guidelines.

There is the danger that these new calculated guidelines represent an over-inflation of effect size in this area of research. Only a limited number of the effects in a limited number of studies within these two journals were used in the calculation. By the nature of the investigation, those findings that were surveyed were generally statistically significant with effect sizes reported for statistically non-significant results in only one study used. As a result of the limited nature of this research, it is at least true to say that these calculations are not as complete or accurate as they ideally would be. Despite this, the new guidelines are very similar to those found in both cognitive and applied psychology, as well as memory research (Morris & Fritz, under review; 2012). Taken together, these results for new effect strength guidelines suggest even more reason to avoid using Cohen's guideline benchmark as a tool for interpreting effect size estimates; they are arbitrary descriptors and, in some cases, simply wrong.

No mention of discrepancies between effect size and statistical significance was mentioned, which may be related to this issue of Cohen's guidelines. All papers that made mention of strength of effect used Cohen's guidelines. However, if Cohen's guidelines are so inaccurate that a small effect, in terms of the new guidelines, is regarded as a medium-large effect using Cohen (1988), then discrepancies may be overlooked. Indeed, Cohen's guidelines in themselves may overestimate the strength of effect. Although there are limits to using any guidelines such as these, they do provide a useful descriptor of the size of the effect that can inform decisions

about practical significance. In addition to this, calculating new benchmarks allows the "direct and explicit comparison" to previous research that Thompson (2007; p. 430) calls for.

Although this research was limited, in that only effects identified in the abstracts were considered, a picture of the present reporting practice can be drawn. There are some positive signs that researchers are considering effect size. However, this consideration has generally been minimal and limited. Although researchers appear to be answering the first two of Kirk's (2001) questions, establishing effect and the size of that effect, they are still neglecting to comment on the practical significance (i.e. the usefulness) of the calculated effect magnitude. On the whole results were reported well in a clear and concise manner, as one would expect of published, peer reviewed papers. Unfortunately there were some cases of very poor reporting, including major findings outlined the abstract not being supported with evidence in the results section. The responsibility to ensure good reporting practice lies with the author, but also arguably the peer reviewers and journal editors that allow research to be published in that form. In addition to the occasional poor reporting of results, there was evidence of misinterpretation of the NHST statistic. Reference was made in one paper to a "marginal" effect being found where the p value calculated was close to, but did not satisfy, the alpha level. This is clearly a misrepresentation or misunderstanding of what this statistic represents. Allowing it to be published suggests that such misconceptions may be widespread, or else the paper was not reviewed thoroughly. Such instances that demonstrate misunderstanding of an almost universally (in terms of educational psychology) used method reiterate questions raised as to the quality of training these researchers have undertaken.

The data suggest that the overreliance on traditional methods (Henson et al., 2010) is still prevalent. Despite some instances which suggest some move toward reform, most researchers relied on NHST along with effect size estimates that were readily available to them. Confidence intervals, and any real engagement with the issue of effect size, have been generally avoided in the papers reviewed. This may reflect a lack of understanding and/or issues of unfamiliarity, proposed by some researchers (e.g., Fritz et al, 2012; Thompson, 2002). Although historically not much attention has been paid to such issues in textbooks (Thompson, 2007), there are now a wealth of resources available to authors. Recent textbooks include those by Cumming (2012), Ellis (2011), Grissom and Kim (2012), and Rosenthal, Rosnow and Rubin (2000). These are comprehensive guides specialising in issues related to effect size estimation, with disciplinary specialism taken to a greater or lesser extent. Also available is a text by Schagen and Elliot (2004) which focuses on effect size issues specifically in relation educational research. In addition to dedicated textbooks, many journals have published research with guides to, as well as the arguments for, effect size reporting. One example is an entire issue of *Educational and Psychological Measurement* (2001; volume 61, issue 4) dedicated to discussion of confidence intervals. Lastly, there are readily available resources online. These include examples of effect size calculators (Ellis, 2009), as well as a comprehensive, and straightforward, guide provided in Fritz, Morris and Richler (2012).

There are many reasons why authors should attend to matters of effect size, from improving the quality of their own studies to matters pertaining to the philosophical considerations of the purpose and ethos of educational psychology research. Whilst

there are resources available, to fully implement reform it may be wise to consider the requirements and quality of training given to under- and post-graduates. It is crucial to provide trainees with good quality, appropriate training to further the development of the discipline. Poor training can lead to poor understanding, demonstrated by poor and incorrect interpretation and reporting of results (as highlighted by Nickerson [2000]). Even assuming the quality of training given is good, the minimal quantitative analysis training required by bodies such as the BPS may simply be inadequate. Although the BPS does require such training from accredited courses (QAA, 2010) typically this translates into just one module on undergraduate courses and sometimes less for postgraduates. The extent to which North American postgraduate psychology courses cover both fundamental and innovative methodology and statistics may also be inadequate (Aiken, West and Millsap, 2008). If the curriculum does not respond to the calls for reform, issues of effect size may well continue to be overlooked by new researchers.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *The American Psychologist, 63 (1),* 32-50.

Alhija, F. N-A., & Levy, A. (2009). Effect size reporting practices in published articles. *Educational and Psychological Measurement, 69 (2),* 245-265.

American Educational Research Association (2001). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35 (6),* 33-40.

American Psychological Association (1994). *Publication Manual of the American Psychological Association (4th ed.).* Washington, DC: American Psychological Association.

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

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

Amos, T., & Kahnemann, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76 (2),* 105-110.

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

Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76,* 159-165.

Bray, J. H. (2010). The future of psychology practice and science. *American Psychologist, 65 (5),* 355-369.

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

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

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

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18 (3),* 230-232.

Ellis, P. D. (2009). Effect size calculators. Retrieved from http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

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

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

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., Schmitt, R. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 73 (1),* 136-143.

Fritz, C. O., Morris, P. E., & Richler, J. (2012). Effect size estimates: Current use, calculations and interpretation. *Journal of Experimental Psychology: General, 141 (1),* 2-18.

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

Harris, D. N. (2008). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis, 31 (1),* 3-29.

Harris, D. N. (2008). Toward policy-relevant benchmarks for interpreting effect sizes: Combining effects with costs. *Educational Evaluation and Policy Analysis, 31 (1),* 3-29.

Henson, R. K. (2006). Effect size measures and meta-analytic thinking in counselling psychology research. *The Counselling Psychologist, 34 (5),* 601-629.

Henson, R. K., Hull, D. M., & Williams, C. S. (2010). Methodology in our education research culture: Toward a stronger collective quantitative proficiency. *Educational Researcher, 39 (3),* 229-240.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62 (2),* 227-240.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis. (2nd ed.).* Thousand Oaks, CA: Sage.

Journal of Educational Psychology (2003). Instructions to authors. *Journal of Educational Psychology, 95 (1),* 201.

Kahnemann, D. (2011). *Thinking, fast and slow.* London: Penguin.

Keselman, H. J., Huberty, C. J, Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., … Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis

of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61 (2),* 213-218.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (ed.), *Handbook of research methods in experimental psychology* (pp. 83-105). Oxford: Blackwell.

Lillienfield, S. O. (2010). Can psychology become a science? *Personality and Individual Differences, 49 (4),* 281-288.

Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F. C., Matthews, D., & Dixon, F. (2008). Evaluating the state of the field: Effect size reporting in gifted education. *The Journal of Experimental Education, 77 (1),* 55–65.

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

Morris, P. E., & Fritz, C. O. (2012, January). *Cumulative measures for interpreting effect sizes*. Paper presented at the London Meeting of the Experimental Psychology Society.

Morris, P. E., & Fritz, C. O. (under review). The reporting of effect sizes in cognitive research.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82 (1),* 3-5.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241-301.

Onwuegbuzie, A. J., Levin, J. R., & Leech, N. L. (2003). Do effect-size measures measure up? A brief assessment. *Learning Disabilities: A Contemporary Journal, 1 (1),* 37-40.

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology, 28 (2),* 151-160.

Peugh, J. L. (2010). A practical guide to multilevel modelling. *Journal of School Psychology, 48 (1),* 85-112.

Quality Assurance Agency (2010). *Subject benchmark statement: Psychology (3rd ed.).* Accessible at www.qaa.ac.uk.

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

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification if knowledge in psychological science*. American Psychologist, 44 (10),* 1276-1284.

Schagen, I., & Elliott, K. (2004). *But what does it mean? The use of effect sizes in educational research.* London: National Foundation for Educational Research.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1(2),* 115-129.

Stanovich, K. E. (2001). *How to think straight about psychology*. Boston, MA: Pearson Education.

Sun, S., Pan, W., & Wang, L. L. (2010). Practices in academic journals in education and psychology. *Journal of Educational Psychology, 102 (4),* 989-1004.

Thompson B. (1999). Journal editorial practices regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review, 11 (2),* 157-169.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25 (2),* 26-30.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31,* 24-31.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44 (5),* 423-432.

Thompson, B. (2008a). Computing and interpreting effect sizes, confidence interval, and confidence intervals for effect sizes. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 246-262). Thousand Oaks, CA: Sage.

Thompson, B. (2008b). http://people.cehd.tamu.edu/~bthompson/index.htm Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology, 51 (4),* 473-481.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54 (8),* 594-604.

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting practices in quantitative teacher education research: One look at evidence cited in the AERA panel report. *Educational Researcher, 37,* 208-216.