The Precision of Effect Size Estimation From Published Psychological Research: Surveying Confidence Intervals

Brand, Andrew; Bradley, Michael T.

Psychological Reports

DOI:
10.1177/0033294115625265

Published: 01/02/2016

Peer reviewed version

Running head: CIs in Psychology

The Precision of Effect Size Estimates from Published Psychological Research – Surveying Confidence Intervals

A. Brand

NWORTH Bangor Clinical Trials Unit, Institute of Medical & Social Care Research, Bangor University

M. T. Bradley

University of New Brunswick

Correspondence concerning this article should be addressed to Michael T. Bradley, Department of Psychology, P.O. Box 5050 Tucker Park Road, University of New Brunswick, Saint John, N.B., Canada, E2L 4L5. Email Bradley@UNB.ca

Abstract

Confidence interval (CI) widths were calculated for reported Cohen's *d* standardized effect sizes and examined in two automated surveys of published psychological literature. The first survey reviewed 1902 articles from *Psychological Science*. The second survey reviewed a total of 5169 articles from across the following four APA journals: *Journal of Abnormal Psychology, Journal of Applied Psychology. Journal of Experimental Psychology: Human Perception and Performance* and *Developmental Psychology.* The median CI width for *d* was greater than 1 in both surveys. Hence CI widths were, as Cohen (1994) speculated, embarrassingly large. Additional exploratory analyses revealed that CI widths varied across psychological research areas and that CI widths were not discernably decreasing over time. The theoretical implications of these findings are discussed along with ways of reducing the CI widths and thus improving precision of effect size estimates.

Key words: accuracy, Confidence Intervals, effect sizes, measurement, precision

The Precision of Effect Size Estimates from Published Psychological Research – Surveying Confidence Intervals

A Confidence interval (CI) provides information about the precision of an effect size estimate by bracketing the unknown true population value of the effect size. Precision, along with accuracy, are hallmarks of science. Hence the integrity of psychological science is undermined if the accuracy and precision of effect size estimates are poor. The concept of accuracy refers to the degree that an estimate conforms to the true population value[1]. Whereas, precision refers to the variability of the effect size estimate and is reflected in the width of the CI. A broad CI width inherently means low precision. Low precision can result from high sample variability or small sample size.

The accuracy and precision of effect size estimates should be assessed because effect size estimates play a key role in the interpretation of research results (see, Kelley & Preacher, 2012). It has been shown that the accuracy of effect size estimates reported in the psychological literature can be severely compromised, for instance, by publication bias (Brand, Bradley, Best, & Stoica, 2008), the averaging and aggregating of data (Brand, Bradley, Best, & Stoica, 2011; Brand & Bradley, 2012), and non-representative sampling (Fielder, 2011). Importantly, however, the precision of effect size estimates in psychology has not been so rigorously assessed.

Calculating and reporting CIs enable researchers to assess the precision of reported effect size estimates. Consequently, Wilkinson and the APA Task Force on Statistical Inference (1999), APA (2001, 2010) recommended that CIs be reported. However, Cumming, *et al.* (2007)

[1] Like, Stallings and Gillmore (1971) and Plant and Turner (2009) we consider accuracy and precision as independent concepts. However, Maxwell, Kelley and Rausch (2008) define accuracy as a function of precision and bias, more specifically the root of the mean square error.

reported in a survey of ten leading international psychology journals that CIs are rarely reported (less than 11% of articles). Cohen (1994) had actually anticipated this by speculating that confidence intervals are rarely reported because they are "embarrassingly large." Given that the sample sizes in psychology studies tend to be small (Marszalek, Barber, Kohlhart, & Holmes, 2011), this seems a reasonable speculation.

As far as we know, no one has actually surveyed the CI width for effect size estimates in psychology. Moreover, quantifying how "embarrassingly large" the CI width for effect size estimates are in psychology is informative. For instance, in the UK during 2014, there was a "depressingly large" amount of rainfall. Though further knowing that during 2014 the UK had 486.8mm of rainfall is informative. Because, for one, it enables us to declare that it was the wettest winter on record. Similarly, knowing the approximate value of the CI width is valuable because it enables us to appreciate the size of the CI width by allowing us to compare the CI width with the magnitude of the effect size estimate. Therefore examining CI width can offer more direct insight into the precision and creditability of psychological research than just surveying sample size, even though CI width is largely determined by sample size.

In this article we conduct two surveys of CI width that were calculated from reported Cohen's d effect size estimates. We initially survey all the articles in *Psychological Science* for the period 2004 to 2012. *Psychological Science* covers the entire spectrum of scientific research in psychology, cognitive, social, developmental, and health psychology, as well as behavioral neuroscience and biopsychology; hence, the research published in *Psychological Science* should be highly representative of psychological research. Additionally, the journal is highly cited with a citation ranking/impact factor that consistently places it in the top ten psychology journals worldwide. To corroborate and extend upon the findings of the first survey, we survey all the

articles 2002 to 2013 from the following APA journals: *Journal of Abnormal Psychology, Journal of Applied Psychology. Journal of Experimental Psychology: Human Perception and Performance Developmental Psychology*. This selection of journals has been previous been considered by Marszalek *et al.* (2011) to be broadly representative of psychology research. Furthermore, this selection of APA journals enables us to examine whether CI widths differ across core areas of psychological research. Finally, we will examine whether CI widths differ across time.

Method

We chose to survey Cohen's *d* effect size estimate because it is a well-known and widely used standardized effect size (Kelley & Rausch 2006), which thus provides a readily comprehendible common metric for CIs. Using correlations (*r*) was considered, however, contrary to APA (2001, 2010) guidelines, degrees of freedom are seldom reported in brackets after the correlation (e.g., *r* (44) = 0.34). The appropriate degrees of freedom would be virtually impossible to extract programmatically.

The surveys were conducted using the statistical package R (R Development Core Team, 2013, http://www.r-project.org/)[2]. First, all the articles in the journals were downloaded. Second, the magnitude of the Cohen's *d* effect size, the associated *t* value and the degrees of freedom were extracted from these articles. Third, the formula: $d = (t*2) / (\sqrt{df})$ was used to calculate the Cohen's *d* effect size in order to determine whether any *d* effect size reported in the text was from a between-subjects design. If the calculated *d* effect size was approximately equal (± 0.01) to the *d* effect size reported in the text, then we assumed that the *d* effect size was from a

[2] The R scripts are available from the author on request.

between-subjects design. This was to allow for rounding error in the reporting of effect sizes and *t* values. Effect sizes that were deemed not from a between-subjects design were omitted because there are no universally agreed upon methods for calculating a standardized effect size and its CIs from repeated measures designs (Morris, 2008). Moreover, any method that calculates the CIs for a repeated measures design would require either the correlation between the repeated measures or the covariance matrix (e.g., Algina & Keselman, 2003). Unfortunately, this information is seldom calculated and never reported.

There is a possibility that effect size estimates from between-subjects design are falsely classified as being not from a between-subject design and are therefore being incorrectly disregarded, since authors may have incorrectly calculated and/or reported the *t* value or effect size estimates. However, we have no reason to believe that excluding *d* effect size estimates from genuine between-subject designs where the reported *d* and *t* value combination was not consistent with a between-subject design will create a systematic bias.

To ensure that the variability and the representativeness of the overall distribution of extracted effect size estimates is not biased; only one effect size estimate is randomly surveyed from articles containing multiple effect size estimates. This importantly enables inferential statistical analyses to be conducted since the independence of data assumption will not be violated.

Finally, the ci.smd() function from the Methods for the Behavioral, Educational, and Social Science (MBESS) R package (Kelley, 2007) using the noncentral t-distribution approach was used to compute the CIs for the reported standardized effect sizes. It was assumed that the sample sizes per groups were equal (or approximately equal if the total sample size was odd)

because information about the sample size per group is not always reported and is also problematic to programmatically extract. However, by assuming the sample size per group is equal, the CIs are computed assuming a best-case scenario, since unequal sample sizes per group result in a greater CI width.

Results

Survey of *Psychological Science*

One thousand and nine hundred and two articles (1902) articles were downloaded from *Psychological Science* for the period 2004 to 2012. Sixty four percent (994) of the extracted *d* effect size etsimates were discarded because they did not correspond to the calculated effect size that assumed a between-subjects design. It should be mentioned that it is highly doubtful whether the discarded effect sizes were all from repeated measures and single sample designs. This is because statistical results are frequently misreported. In a survey of 281 articles from the psychological literature, Bakker and Wicherts (2011) found that around 18% of statistical results were incorrectly reported. After randomly selecting one *d* effect size estimate per article, a final sample of 149 effect size estimates was obtained.

Figure 1 – About Here

The distribution of effect sizes is positively skewed, the median is 0.66 (inter-quartile range [IQR] = 0.47-0.91) Given that statistically significant results ($p < 0.05$) are more likely to be published (Francis, 2014, Sterling, 1959) and that the sample size in studies are typically less than 100, it is not surprising that the majority of the reported effect sizes were equal or greater to Cohen's (1977) medium effect size benchmark of 0.50.

Figure 2 – About Here

The distribution of the widths of the 95% CI was positively skewed, the median was 1.07 (inter-quartile range [IQR] = 0.85-1.31). Notably, both the median and the lowest IQR value were larger than Cohen's (1977) large effect size benchmark of 0.80.

Figure 3 – About Here

The distribution of the 95% CI widths as a percentage of effect size was positively skewed, the median was 164% (inter-quartile range [IQR] = 120%-195%). Just to give some perspective, 15% (22 out of 149) of the CI widths were lower than the reported effect size, whereas 85% (126 out of 149) widths were greater and 22% (33 out of 149) of the CI widths were twice as large as the reported effect size estimate.

Survey of APA Journals (*Journal of Abnormal Psychology, Journal of Applied Psychology. Journal of Experimental Psychology: Human Perception* and *Performance Developmental Psychology)*

In total 5169 articles were surveyed from the four APA journals for the period 2002 to 2013. Similarly, to the survey of *Psychological Science*, a high percentage (72%, 1249 out of 1738) of the extracted *d* effect size estimates were discarded because they did not correspond to the calculated effect size that assumed a between-subjects design. A final a sample of 141 effect size estimates was obtained, after one *d* effect size estimate per article was randomly sampled. As discussed previously, we suspect that a marked percentage (approximately 20% or higher) of the discarded effect size estimates were actually misreported effect sizes from between-subjects

designs. To verify this conjecture we randomly selected, thirty effect estimates that were classified as being not from a between-subjects design and manually classified them. The results confirm our conjecture, 67% (20) were from a repeated measures design, 7% (2) were from a single sample design and 27% (8) were from a between-subjects design.

Figure 4 – About Here

The distribution of effect sizes is positively skewed, the median was 0.57 (inter-quartile range [IQR] = 0.30-0.89)

Figure 5 – About Here

The distribution of the widths of the 95% CI was positively skewed, the median was 1.03 (inter-quartile range [IQR] = 0.74-1.35).

Figure 6 – About Here

The distribution of the 95% CI widths as a percentage of effect size was positively skewed, the median was 163% (inter-quartile range [IQR] = 122%-228%). Of the distribution, 17% (24 out of 141) of the CI widths were lower than the reported effect size, whereas 82% (115 out of 141) widths were greater and 30% (43 out of 141) of the CI widths were twice as large as the reported effect size estimate.

Overall the results from the APA journals survey are inline with the results from the survey of *Psychological Science*. Next we examined the CI width of the effect size estimates from APA journals from different psychological research areas (see Figure 7).

Figure 7 – About Here

A Kruskal Wallis test revealed a statistically significant effect of journal on CI width ($H(3)=16.62$, $p < .001$). Post-hoc tests using Mann-Whitney U tests showed that there was a statistically significant differences between the CI widths from: *Developmental Psychology* and *Journal of Applied Psychology* ($U = 1147$, $p < .001$, Hodges–Lehmann estimator = 0.36, 95% CI [0.16, 0.59]); *Journal of Abnormal Psychology* and *Journal of Applied Psychology* ($U = 1087$, $p = .009$, Hodges–Lehmann estimator = 0.23, 95% CI [0.07,0.39]); *Journal of Applied Psychology* and *Journal of Experimental Psychology: Human Perception and Performance* ($U = 248$, $p = .005$, Hodges–Lehmann estimator = 0.79, 95% CI [0.21, 1.34]). The post-hoc tests also revealed that there were no statistically significant differences between the CI widths from: *Developmental Psychology* and *Journal of Abnormal Psychology* ($U = 1352$, $p = .206$, Hodges–Lehmann estimator = 0.14, 95% CI [-0.06, 0.37]); *Journal of Experimental Psychology: Human Perception and Performance* and *Developmental Psychology* ($U = 336$, $p = .166$, Hodges–Lehmann estimator = 0.37, 95% CI [-0.15, 0.95]); *Journal of Experimental Psychology: Human Perception and Performance* and *Journal of Abnormal Psychology* ($U = 371$, $p = .054$, Hodges–Lehmann estimator = 0.54, 95% CI [-0.01, 1.12]).

Finally, we examined whether the CI width from effect size estimates vary across time by collating the width of the CIs from the Psychological Science and the APA journals for the period 2005 to 2012. Hopefully, as time progresses the CI width of the effect size estimates would decrease but this was found not to be the case (see Figure 8)

Figure 8 – About Here

A robust linear model using MM-type regression estimator (see, Yohai, 1987) was fitted to the data. It revealed that there was no statistically significant linear relationship between publication year and CI width, b = -0.00, t (255) = -0.38, p = .704, 95% CI [-0.03, 0.02].

In summary, these additional exploratory analyses showed that CI widths varied across psychological research areas and that CI widths were not discernably decreasing over time. Hence these results reflect the findings from the survey of sample sizes conducted by Marszalek *et al.* (2011).

Discussion

This empirical sample supports Cohen's point, that confidence intervals are typically very large. The implications of this are striking. Especially if we consider that 83% of the effect size estimates sampled had CI width that were larger than the reported effect size estimate and 26% were twice as large as the reported effect size estimates. Would we want to rely on a weather forecaster that predicts with 95% confidence that the temperature will be 20ºC ± 10ºC (i.e., the estimate is 20ºC and 95% CI width for the estimate is 20ºC)? Likewise, if a CI width is the same size or larger than the effect size estimate the creditability of the effect size is also highly questionable. It would be completely possible for a researcher to conduct a direct replication of a study but obtain a substantially different effect size estimate[3]. For example, Researcher A conducts an experiment and obtains a Cohen's *d* effect size estimate of 0.60. Researcher B conducts a direct replication of the original experiment with the same sample size, and obtains a

[3] Defining a successful replication as obtaining an effect size estimate that lies within the CI width of the original experiment, though plausible, is unworkable because studies that have small sample sizes would be the most replicable.

Cohen's *d* effect size estimate of 0.30. This discrepancy between the effect size estimates undermines our faith and confidence in the findings of both experiments. Hence, the results of the CI survey strongly support the view that there is a, so-called, "crisis of confidence in psychological science" (Pasher & Wagenmakers, 2012). More importantly, the findings of the CI survey provide us with a new appreciation of the scale and extent of this crisis and in doing so highlights the need for CIs to be substantially reduced in future studies.

Fortunately, there are several ways of reducing confidence interval width. Cohen (1990) suggested employing data from past studies to evaluate results from current and future studies. His method involves reducing the level of confidence to 80%. Usually, only the CIs for 99%, 95%, 90% confidence are reported (Altman, Machin, Bryant & Gardner, 2000). However, a reduction in confidence from 95% to 80% has a marked effect. The median for the 80% CI width for our survey reduces the median CI width by 35%, to 0.72. On the one hand, estimates that fit in this reduced interval have passed a more rigorous test, but the cost or tradeoff is less certainty that a given estimate does belong in the interval. Thus, the loss of confidence about fit in the interval makes the value in the reduction in the breadth of CIs an unappealing proposition. Another approach suggested by Cohen (1994) that shrinks CI width is to reduce invalid and unreliable variance in the dependent measure. This is a suggestion that in practice is done to some degree through normal measurement design. At a certain point, however, measurement refinements may lead to diminishing returns or are not possible given the state of knowledge in a given area. Also decreasing the variance in the dependent measure will not decrease the CI width of a standardized effect size estimate such as Cohen's d.

The most obvious approach is to simply increase the sample size (Smithson, 2003). However, to obtain a desirable CI width, reflecting Cohen's small or medium size benchmarks

requires what have been heretofore considered prohibitively large sample sizes. To illustrate this we have calculated the sample sizes required to obtain a CI width of 0.2 and 0.5 for Cohen's (1977) small, medium and large effect size benchmarks using the accuracy in parameter estimation (AIPE) approach developed by Kelly and Rausch (2006) see Table 1[4]. The large sample sizes required for investigating small effects may be only attainable by multi-centre studies (Maxwell, 2004) or web-based studies (Birnbaum 2004) where appropriate[5]. Hence, future psychological research would have to either be more collaborative and/or employ web-based technologies to achieve effect size estimates with credible precision.

Table 1 - About Here

Increasing sample size and therefore increasing statistical power will reduce both the CI width and improve the accuracy of published effect size estimates assuming the bias to predominantly publish statistically significant findings ($p < 0.05$) persists[6]. However, other factors such as measurement error and the inappropriate statistical treatment of data can compromise the accuracy of effect size estimates, even when sample sizes are increased and CI widths are reduced. Thus, even if the CI width of effect size estimates can be sufficiently reduced by increasing sample, the creditability of psychological research may still be compromised by the accuracy of the effect size estimates. There is therefore, unfortunately, no single solution to achieving both precise and accurate effect sizes estimates.

[4] Beal (1989) and Liu (2009, 2012) also provide methods for determining sample sizes to achieve a desired CI width.

[5] Note for instance, web-based studies would not be suitable for perceptual experiments where the viewing conditions need to be strictly controlled.

[6] Note there is a statistical method proposed by Bradley and Stoica (2004) and further endorsed by Bradley and Brand (2013) that can correct for the effect of publication bias on published effect size estimates.

Ideally, to make the present surveys more complete, we would like to have calculated and surveyed the CIs for the two group repeated measure designs. Unfortunately, there are no universally agreed upon methods for calculating a standardized effect size and its CI for a repeated measures design. Furthermore, calculating CIs for a repeated measures design would require access to the raw data and this is logistically problematic because it would involve contacting 2243 authors. Moreover, requests for data in psychology are seldom fulfilled (Wicherts, 2006).

Given that we have only calculated and surveyed the CIs for the Cohen's *d* effect size, an effect size that assumes a between-subjects design, we must be careful not to over-generalize. However, even though the CIs for effect size estimates from a repeated measure designs will be narrower than the CIs for effect size estimates from between-subjects designs (Cummings, 2012), the present findings would suggest that CIs for a repeated measures design will also be unacceptably wide, simply because of the sheer magnitude of the CI width observed for the between-subjects design.

The automated approach we used to survey the journal articles allowed us to conduct two large-scale surveys efficiently in a time effective manner. One unforeseen consequence of the automated approach we employed is that an expectedly high percentage of effect size estimates were discarded because they were not classified as being from a between-subject design. It is therefore likely that numerous effect size estimates from between-subjects design were discarded. This may appear problematic. However, the exacting nature of the automated approach we employed (using the *t* value and the degrees of freedom) to determine whether a design is a between-subjects design also acts as a quality control. This is because it filters out studies where the *t* value and effect size has been incorrectly calculated or reported.

One should finally bear in mind that the surveys examined the reported statistical results and not the raw data, since a large-scale survey of raw data would be impractical. Therefore an implicit assumption of the surveys conducted is that statistical test assumptions of the t-test (e.g., homogeneity of variance and normality of residuals) have not been violated and statistics (e.g., *t* value, *d* effect size) have been calculated and reported correctly. These assumptions, one would hope, are met in the majority of publish peer-reviewed research.

In summary, the results from the automated surveys of CI widths quantify and highlight the poor precision of effect size estimates from published psychological research. Effect size estimate imprecision was prevalent in all the areas of psychological research examined. Furthermore, the surveyed CI widths have not discernably reduced over time hence no improvement in the precision of effect size estimates was observed. Although there are methods for reducing CI widths and improving the precision of effect size estimates, as discussed, such methods are problematic. Reducing the breadth of an interval inherently reduces the confidence level or the very quality the calculation approach is meant to capture, and other methods are non-trivial to implement requiring a fundamental change to how psychological research is conducted.

Acknowledgement

We would like to thank the American Psychological Association (APA) for their cooperation.

References

Algina, J., & Keselman, H. J. (2003) Approximate confidence intervals for effect sizes. Educational and Psychological Measurement, *63*, 537-553.

Altman, D. G., Machin, D., Bryant, T. N., & Gardner, M. J. (Eds.) (2000) Statistics with Confidence: Confidence Intervals and Statistical Guidelines (2nd ed.) London: British Medical Journal.

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

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

Bakker, M., & Wicherts, J. M. (2011) The (mis)reporting of statistical results in psychology journals. Behavior Research Methods, *43*, 666-678.

Beal, S. L. (1989) Sample size determination for confidence intervals on the population mean and on the difference between two population means. Biometrics, *45*, 969-977.

Birnbaum, M. H. (2004) Human research and data collection via the internet. Annual Review of Psychology, *55*, 803-832.

Bradley, M. T. & Stoica G. (2004) Diagnosing estimate distortion due to significance testing in literature on detection of deception. Perceptual and Motor Skills, *98*, 827–839.

Bradley, M. T. & Brand, A. (2013) Sweeping recommendations regarding effect size and

sample size can miss important nuances: A comment on: A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? Theory and Psychology. 23, 6

Brand, A., Bradley, M. T., Best, L., & Stoica, G. (2008) Accuracy of effect size estimates from published psychological research. Perceptual and Motor Skills, *106*, 645-649.

Brand, A., Bradley, M. T., Best L. A., & Stoica, G. (2011) Multiple trials may yield exaggerated effect size estimates. The Journal of General Psychology, *138*, 1-11.

Brand, A., & Bradley, M.T. (2012) More Voodoo Correlations: When averaged based measures inflate Correlations. The Journal of General Psychology, *139*, 260-272.

Cohen, J. (1977) Statistical Power Analysis for the Behavioural Sciences. New York: Academic Press.

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

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

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*, 230-232.

Cumming, G. (2012) Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. New York: Routledge.

Eisenhart, C, (1968) Expression of the uncertainties of final results: Clear statements of the uncertainties of reported values are needed for their critical evaluation. Science, *160*, 1201-1204.

Fiedler, K. (2011) Voodoo correlations are everywhere—not only in neuroscience. Perspectives on Psychological Science, *6*, 163–171.

Francis, G. (2014) The frequency of excess success for articles in Psychological Science. Psychonomic bulletin & review, *21*, 1180-1187.

Kelley, K. (2007) Methods for the Behavioral, Educational, and Educational Sciences: An R package. Behavior Research Methods, *39*, 979-984.

Kelley, K., & Preacher, K. J. (2012) On effect size. Psychological methods, *17*, 137-152.

Kelley, K., & Rausch, J. R. (2006) Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. Psychological Methods, *11*, 363-385.

Liu, X. (2009) Sample size and the width of confidence interval for mean differences. British Journal of Mathematical and Statistical Psychology, 62, 201-215.

Liu, X. (2012) Implications of statistical power for confidence intervals. British Journal of Mathematical and Statistical Psychology, *65*, 427-437.

Marszalek, J.M., Barber, C., Kohlhart, J., & Holmes, C.B. (2011) Sample size in Psychological research over the past 30 years. Perceptual and Motor Skills, *112*, 331-348.

Maxwell, S. E. (2004) The persistence of underpowered studies in psychological research: Causes, consequences, and Remedies. Psychological Methods, *9*, 147-163.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008) Sample size planning for statistical power and accuracy in parameter estimation. Annual Review of Psychology, *59*, 537–563.

Morris, S. B. (2008) Estimating effect sizes from pretest-posttest-control group designs. Organizational Research Methods, *11*, 364-386.

Pashler, H. & Wagenmakers, E.-J. (2012) Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? Perspectives on Psychological Science, *7*, 528-530.

Plant, R. R., & Turner, G. (2009) Millisecond precision psychological research in a world of c ommodity computers: New hardware, new problems? Behavior Research Methods, *41*, 598-614.

R Development Core Team (2013) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.r-project.org.

Smithson, M. (2003) Confidence intervals. Quantitative Applications in the Social Sciences Series, No. 140. Thousand Oaks, CA: Sage.

Stallings, W.M. & Gillmore, G.M. (1971) A Note on "Accuracy" and "Precision", Journal of Educational Measurement, *8*, 127-129.

Sterling, T. D. (1959) Publication decisions and their possible effects on inferences drawn from tests of significance or vice versa. Journal of the American Statistical Association, *54*, 30-34.

Wicherts, J.M., Borsboom, D., Kats, J., & Molenaar, D. (2006) The poor availability of psychological research data for reanalysis. American Psychologist, *61*, 726-728.

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

Yohai, V. J. (1987) High breakdown-point and high efficiency robust estimates for regression. The Annals of Statistics, *15*, 642-656.

Figure 1: The random selection of 149 Cohen's *d* standardized effect size estimates from the Psychological Science articles that met the requirement for a between-subjects design

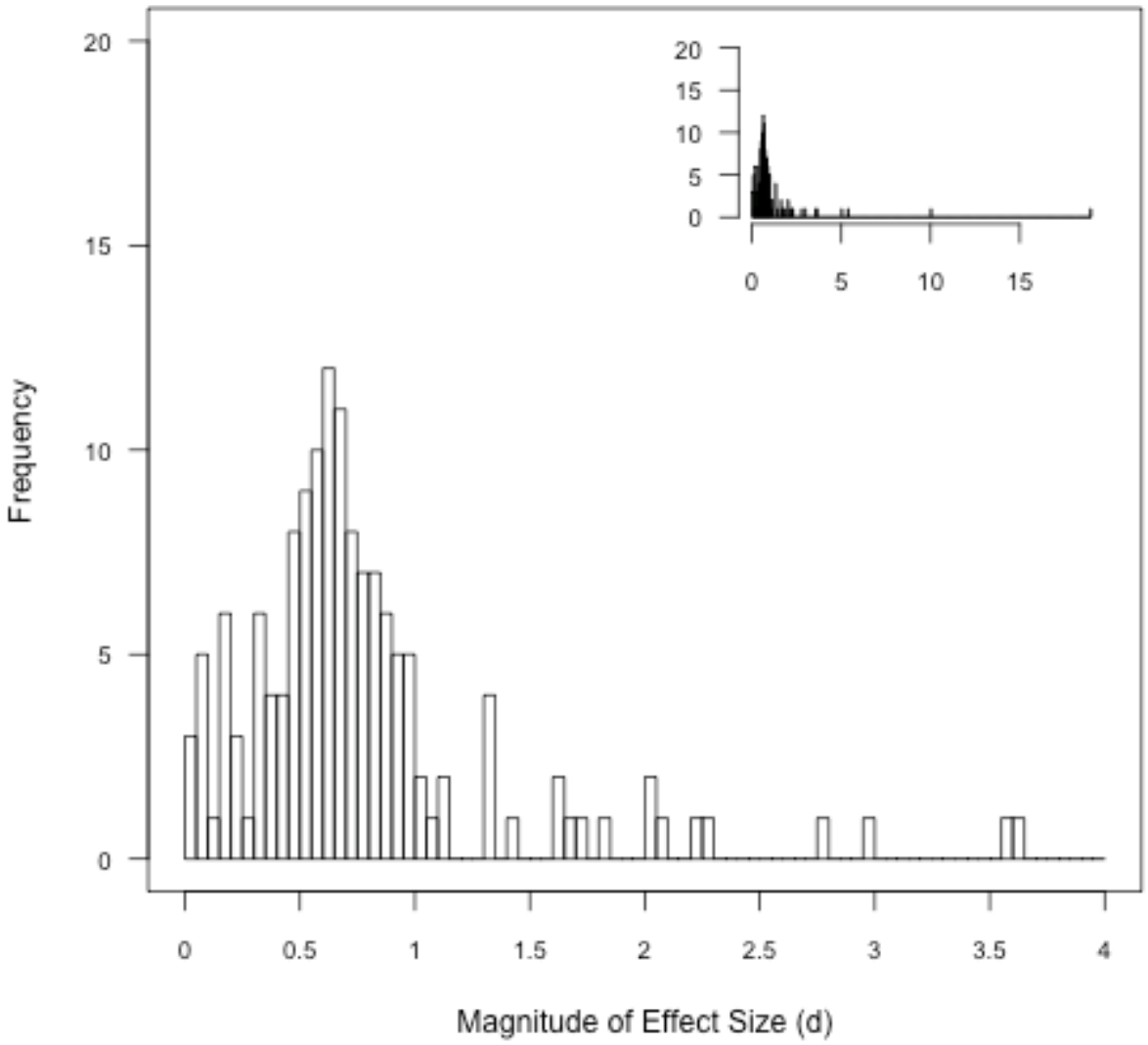


Figure 2: The Width of the 95% CI for the 149 Cohen's *d* standardized effect size estimates from the Psychological Science articles

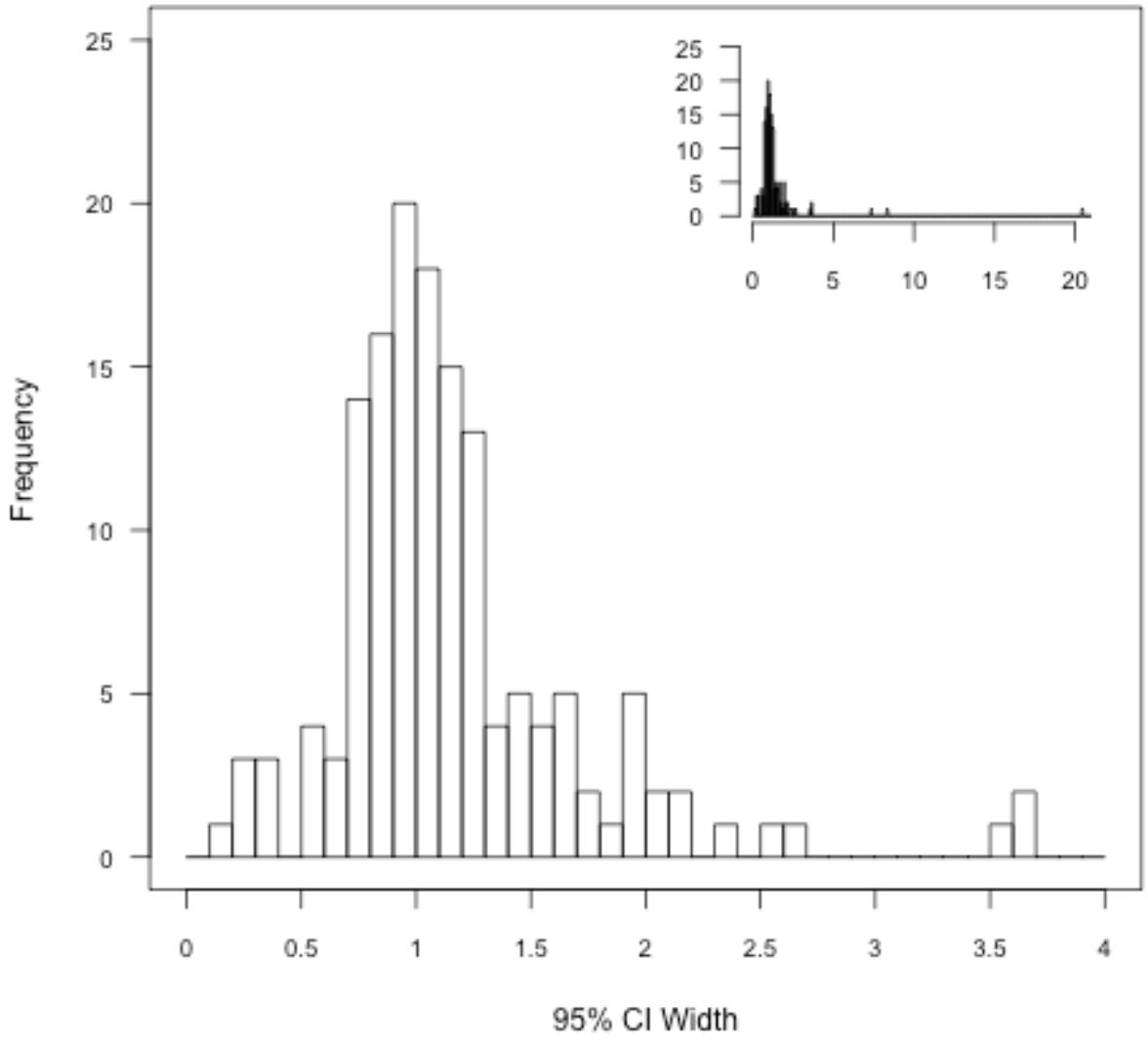


Figure 3: The Width of the 95% CI as a percentage of the reported Cohen's *d* standardized effect size estimate from the Psychological Science articles

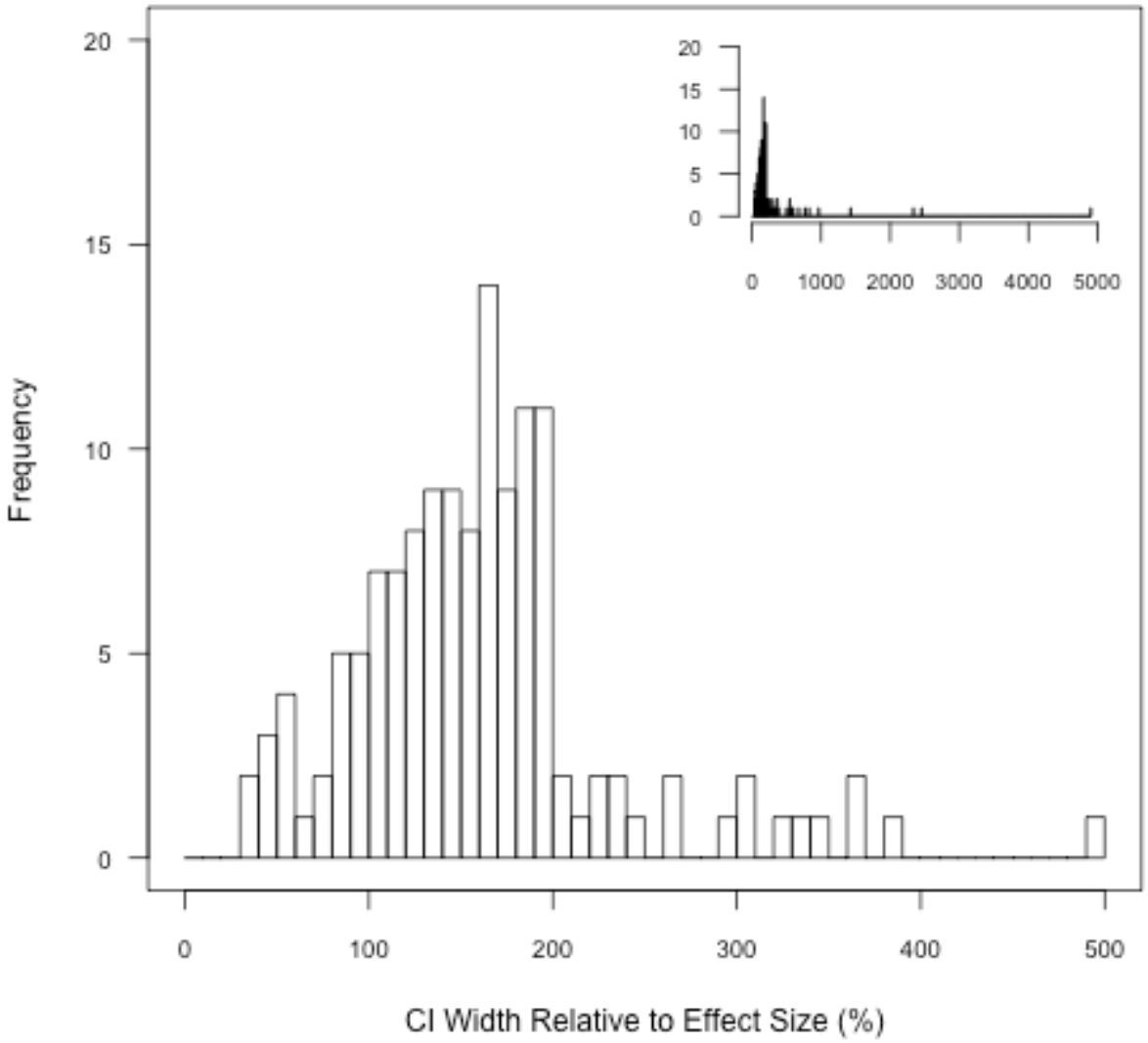


Figure 4: The random selection of 141 Cohen's *d* standardized effect size estimates from the APA journal articles that met the requirement for a between-subjects design

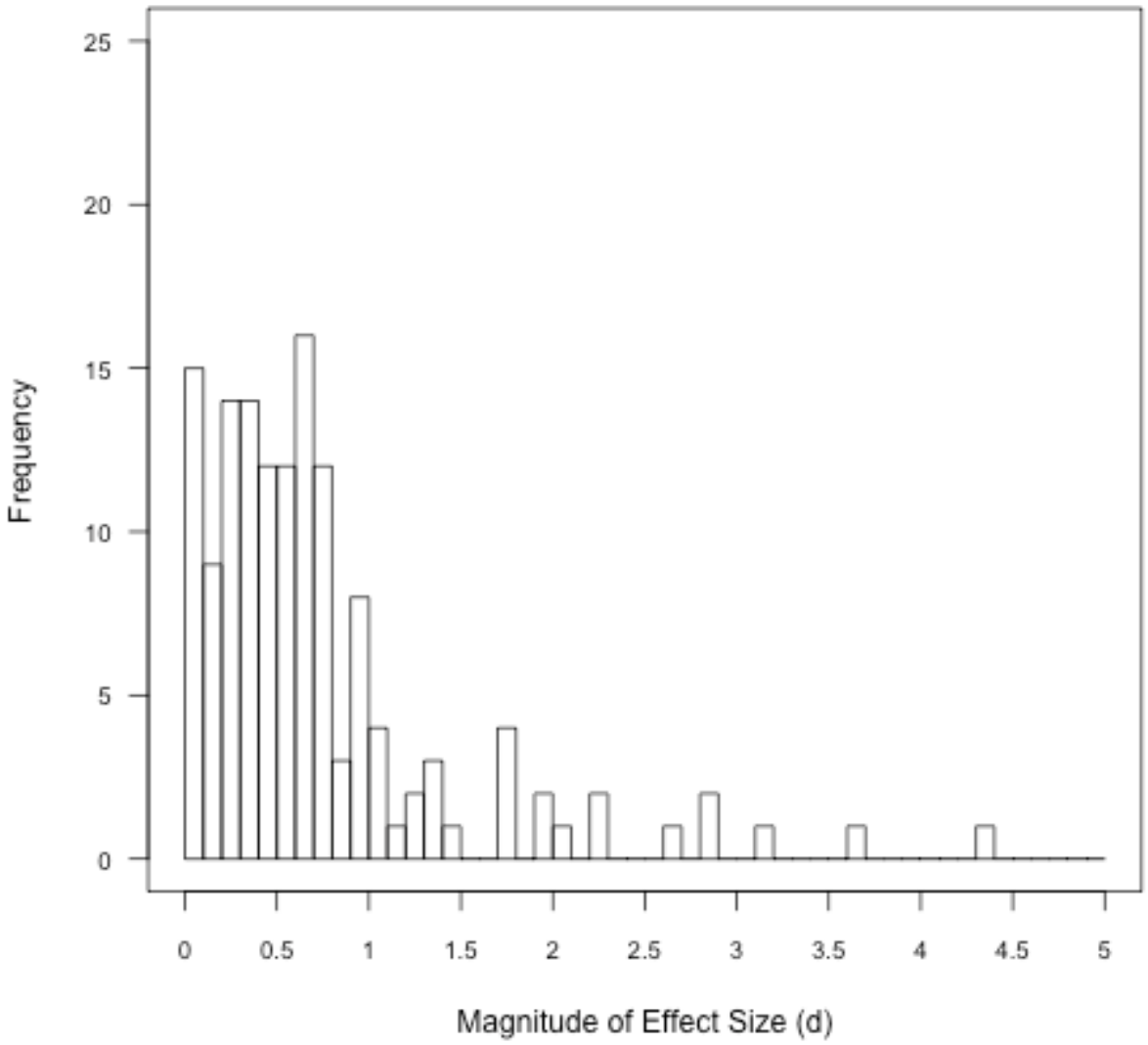


Figure 5: The Width of the 95% CI for the 141 Cohen's *d* standardized effect size estimates from the APA journal articles

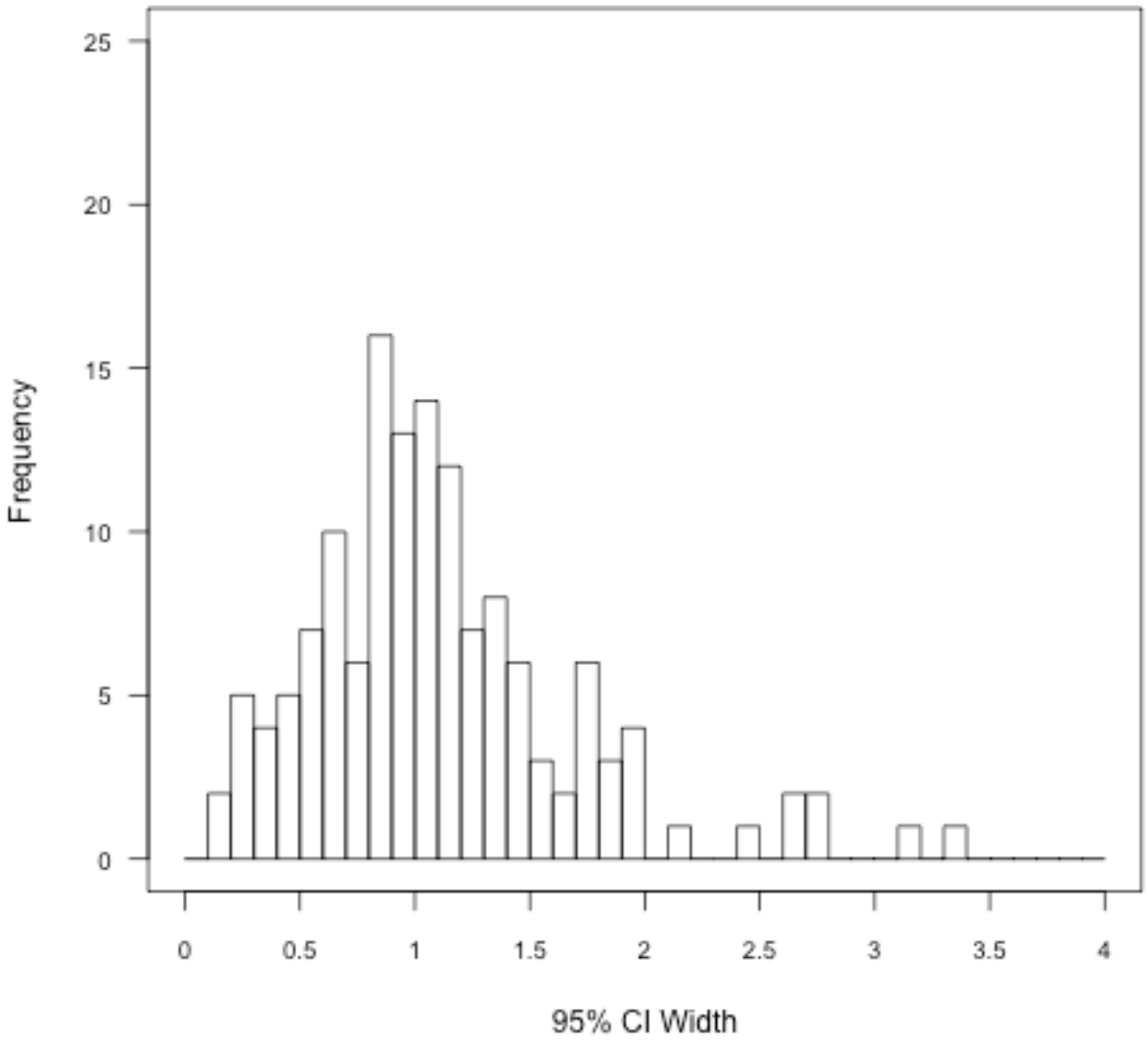


Figure 6: The Width of the 95% CI as a percentage of the reported Cohen's *d* standardized effect size estimate from the APA journal articles

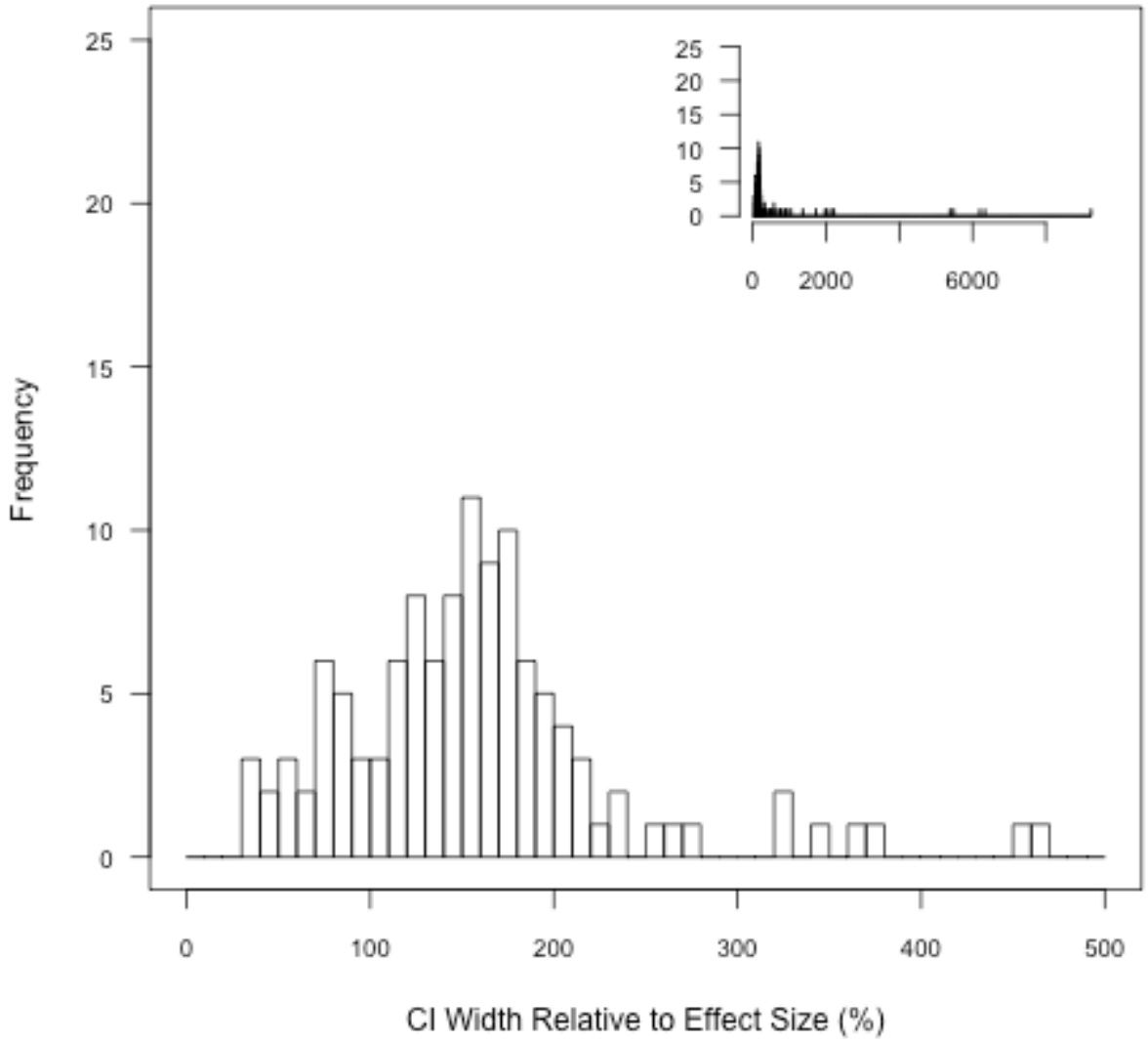


Figure 7: Boxplots of CI Width for the four APA Journals

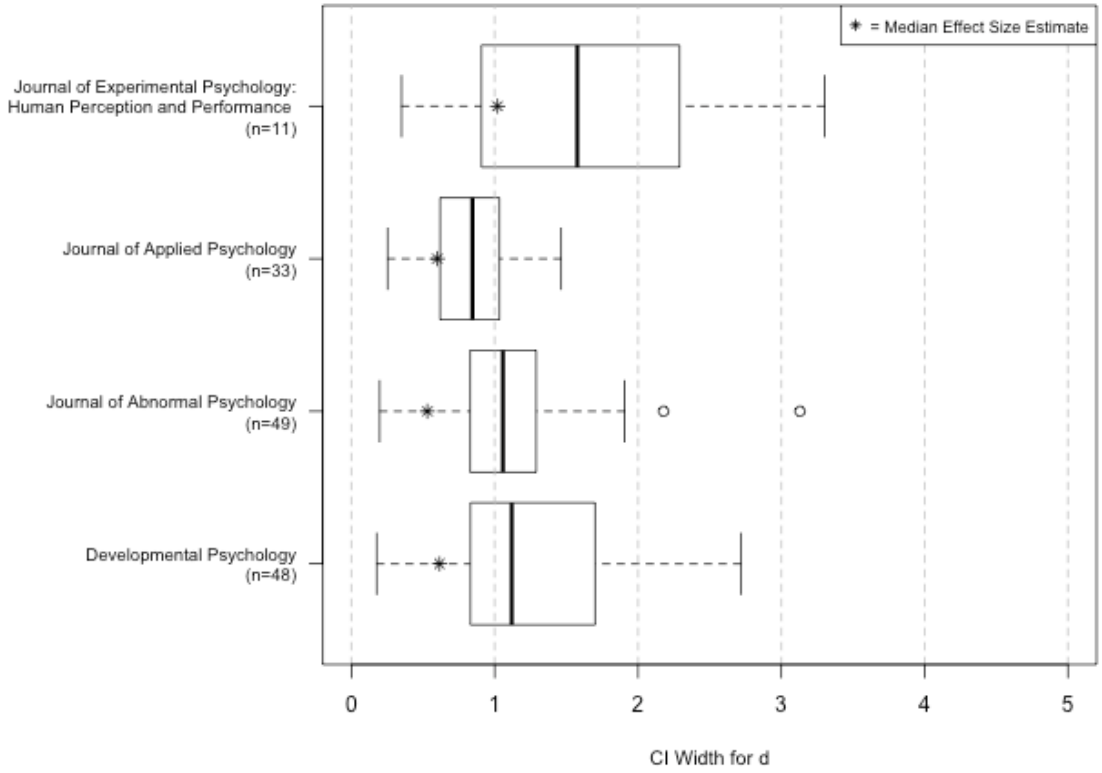
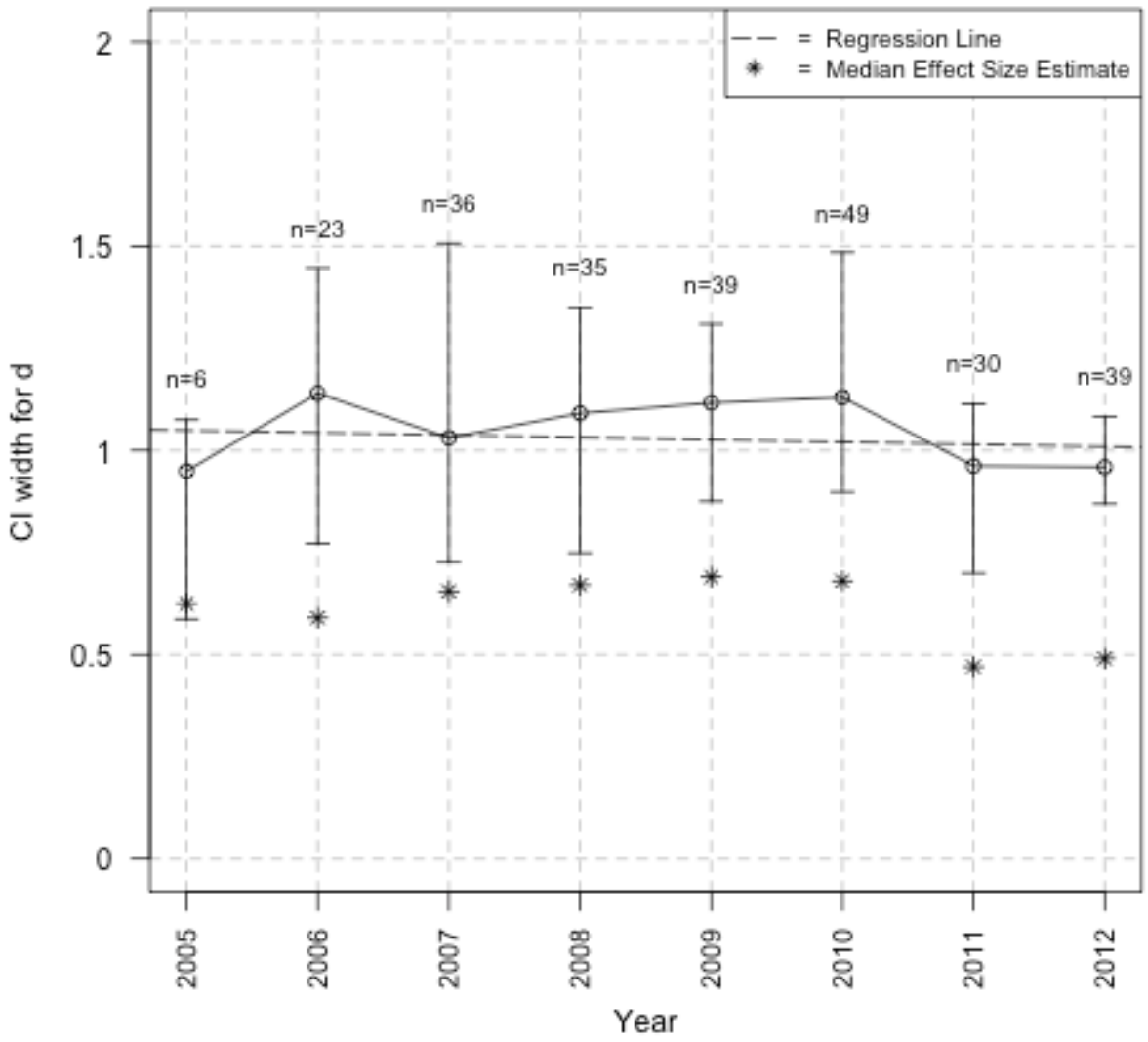


Figure 8: Plot of the Median CI Widths over time



Note that the lower error bar represents the first quartile and the upper error bar represents the third quartile.

Table 1: Sample Sizes required to obtain 95% CI Width of 0.20, 0.50 and 0.80 as a function of the True Effect Size

		True Effect Sizes		
		$d = 0.20$	$d = 0.50$	$d = 0.80$
CI width	0.20	1554	1604	1688
	0.50	252	262	278