Effect Sizes Over Time

Felix Singleton Thorn

## Abstract

This article uses mixed effects meta-regression to estimate the change in average effect sizes in psychological research using a database of over 130,000 effect size estimates from over 9,000 articles published in 5 APA journals from 1985 to 2013 {Nuijten, 2015 #550}. The results of this analysis suggest that the average effect size reported in psychological research is decreasing by -0.004 (95% CI [-0.005, -0.004]) Fisher Z score units per year. This represents a substantial decrease of -0.1 correlation coefficient units from 1985 to 2013. Several exploratory analyses were also performed to investigate whether this change is also seen in subsets of reported effects that may be more likely to represent the effects of substantive interest; the first reported effect and the largest effect in each paper. Analysing just the first detected effect size in each paper showed broad agreeement with the main analysis (a -0.003 estimated yearly change, 95% CI [-0.004, -0.002]), however, looking at the largest effect reported in each paper showed a slight increase in effect sizes over time (an estimated yearly change of 0.0015 (95% CI [0.0002, 0.0028]) ). Together these results suggest that there has been a small decrease in the average effect sizes reported in psychology over time, although the degree to which this decrease is reflective of a decrease in the size of the focal or main effects under study in psychology remains an open question.

## Introduction

An important question in understanding the history of psychological science is whether and how the magnitudes of the effects under study have changed over time. Are we studying smaller effects over time, having already studied the “low hanging fruit” {Baumeister, 2016 #1022}? Or could methodological reforms and the increasing awareness of the importance of measurement in research design lead to increased effect sizes {Nelson, 2018 #750;Greenland, 2017 #713;Loken, 2017 #164}? This paper uses an extensive database of over 130,000 effect size estimates from almost 10,000 articles published in 5 APA journals from 1985 to 2013 collected as part of {Nuijten, 2015 #550} to examine how effect sizes have changed in psychological research over this period.

The question of how effect sizes have changed over time is both of intrinsic interest and has important implications for understanding the results of surveys of the statistical power of psychological research (e.g., {Cohen, 1962 #487;Rosnow, 1989 #37;Szucs, 2017 #25}). Most of the efforts to estimate the statistical power of psychological research have used Cohen’s effect size benchmarks {Cohen, 1988 #562}, and as such any comparison of these studies over time assumes that effect sizes have been stable. According to our recent meta-analysis 46 power surveys including over 8,000 individual studies published from 1932-2014 [cite meta-analysis], the average statistical power of psychological research is .23 95% CI [.17, .29] for “small” effect sizes (effect sizes equivalent to r = .1 or Cohen’s d = .2), .62 95% CI [.55, .69] to detect medium effects (r = .3 or Cohen’s d = 0.5)), and .80, 95% CI [.68, .92] to detect large effects (effects equivalent to 0.8 Cohen’s d or r = .5) [cite meta-analysis]. This same analysis also suggests that there has been little-to-no change in the average statistical power of research conducted in psychology to detect these effect size benchmarks over time. However, in order to know whether and how the average statistical power of psychological research has changed, it is necessary to know whether the effect sizes under study in psychological research have remained stable.

One previous study looked at effect sizes reported as odds ratios (including odds ratios, risk ratios and hazard ratios) in abstracts in the PubMed database {Monsarrat, 2017 #1060}. Monsarrat and Vergnes (2018) showed that effect sizes reported in this literature are, on average, decreasing over time. However, it is unclear whether this trend is also seen in the psychology literature. A small number of previous studies have extracted effect size benchmarks from psychological research in domains as varied as social {Richard, 2003 #603}, management {Paterson, 2015 #817; Bosco, 2015 #157} and clinical psychology {Haase, 1982 #516}. Estimates of the average effect size in various subfields range from an mean correlation coefficient of .21 in social psychology meta-analyses {Richard, 2003 #603}, to a mean effect of 0.94 Cohen’s d (equivalent to r = .42) seen across statistical tests reported in recent cognitive neuroscience, psychology and psychiatry articles from high impact journals {Szucs, 2017 #25}. However, we have not identified any studies which have adequate statistical precision to make strong inferences about the size or direction of the change in effect sizes over time. We know of just one study which attempted to examine the degree and direction of change in average effect sizes over time. {Paterson, 2015 #817} found a small negative correlation (r = -.05, 95% CI [-0.121, 0.02]) between the reported magnitude of correlations and their year of publication in 776 meta-analytic conclusions from meta-analyses in management psychology. The current analysis allows us to examine the average effect size reported in psychological research, and allows us to precisely estimate the change in effect sizes over time across fields of psychological research.

### Method

This analysis uses the dataset developed in {Nuijten, 2015 #550}, a study examining the number of errors in statistical tests reported in psychology articles (all data from <https://osf.io/gdr4q/>). This dataset includes 258105 statistical test results from 16695 articles. Nuijten et al. (2018) extracted this database using regular expressions, exploiting the APA style guide’s strict rules for reporting statistical test results and to extract all chi square, t tests, F tests, Z tests, and correlations reported in APA style from articles published in eight major psychology journals from 1985 to 2013.

The current paper only uses a subset of these studies, those reporting in journals for which results were available going back to 1985. This analysis therefore includes data from five psychology journals chosen to be representative of the main subdisciplines of psychology research; Journal of Applied Psychology (JAP; Applied Psychology), Journal of Consulting and Clinical Psychology (JCCP; Clinical Psychology) , Developmental Psychology (DP; Developmental Psychology), Journal of Experimental Psychology: General (JEPG; Experimental Psychology), and Journal of Personality and Social Psychology (JPSP; Social Psychology). The excluded Journals (PLOS, Psychological Science, Frontiers in Psychology) have only began to be published in the last 15 years. The included subset includes a total of 200763 statistical test results from a total of 11825 articles published from 1985 to 2013.

### Effect size extraction and conversion

All test statistics were converted to Fisher Z transformed correlation coefficients (henceforth ) following {Open Science Collaboration, 2015 #611} for visualization and analysis, see Supplementary Materials 1 for detailed explanations of the transformations used. Negative effect sizes (i.e., negative correlations) were set to be positive for analysis and visualization. Standard errors were estimated as , taking the degrees of freedom for the denominator minus two as n (or just the degrees of freedom minus two in the case of correlations and t tests). Because typical APA notation for z tests does not report the included sample size in a standardized format (and therefore this information was not available in this dataset), Z scores were excluded from analyses (n = 7539). As it was not possible to derive valid standard errors for F statistics with denominator degrees of freedom above 1 (n = 19713) or for Chi square statistics (n = 21855), these analyses were excluded from the multilevel meta-analysis (see Figure 1 for histograms comparing effect sizes included in each analysis and those derived from the entire sample). An additional subset of results were excluded from all analyses as they produced standard errors or Z transformed correlation coefficients which could not be estimated or were infinite (n = 767, e.g., studies which reported impossible test statistics such as “F(0, 55) = 5.71, p < .05” or “r(66) = 5.42, p < .001” and studies with denominator degrees of freedom below 6). This left a total of 132086 effect size estimates with valid standard errors from 9472 articles were extracted and are included in the meta-regression analyses below.

## Analysis

Multilevel meta-regression was performed to examine the relationship between year of publication and reported effect sizes.

The multilevel-meta-regression includes random effects for individual tests (), articles () and journals (), and includes year of publication of each article as a fixed effect (). This analysis was performed using nlme {Pinheiro, 2018 #1027} in R version 3.5.1 {R Development Core Team, 2018 #314} using restricted maximum likelihood estimation.

As a check on whether the exclusion of F statistics with of greater than one and analyses is likely to change the results, this multilevel model was reperformed including all of the statistical tests for which correlations could be estimated and estimating standard errors as for F tests and for tests. This analysis includes a total of n = 170556 effects after excluding all invalid results (i.e., analyses where it was not possible to estimate a standard error or effect size using the above methods due to issues such as impossible test statistic values that were likely typos). This analysis led to estimates of the change in effect sizes over time which are practically identical to those presented below (i.e., which differed by < 0.002).

Robustness checks were performed to assess the robustness of these results to model specification changes. Including random slopes by journal, fixed effects for test statistic type, allowing the change over time to vary by test statistic type, or not accounting for the variance of each effect size estimates did not lead to any substantial difference in the results of this analysis. See Supplementary Materials 2 for a more in-depth discussion of each of these models.

#### Exploratory analyses Accounting for ‘peripheral’ tests

Due to the large sample size included in this analysis and the extremely small expected effects (and hence, large sample sizes required), it was not feasible to manually label which statistical tests included were, for example, manipulation tests or randomization tests, and some of the tests reported and included in the current analysis are certainly not tests of the substantive hypotheses of each included paper. This means that any observed change in effect sizes could be driven by changes in reporting practices, such as a changing number of reported manipulation or randomization checks over time.

In order to account for this issue, I performed two additional exploratory analyses to examine whether the trends seen across all reported statistical tests are also seen in subsets of results that may be more likely to represent the “main” effects under study. In order to do so multilevel mixed effects meta-regression was performed looking just at (a) the first statistical test reported in each paper and (b) the largest effect size reported in each paper. For (b), in 84 cases where there were ties within papers for the largest effect size, the first of the two equal outcome sizes was used.

These multilevel-meta-regressions included random effects for individual articles () and journal (), and included year of publication of each article as a fixed effect (). These analyses did not include random effects for each statistical test, as each article only provides a single effect size. Valid standard errors were calculable for effect sizes from a total of 9472 articles for analyses (a) and (b), all of which were included in these analyses. These analyses were performed using the R package metafor {Viechtbauer, 2010 #796} using restricted maximum likelihood estimation.

### Deviations from preregistration

All analyses were tested and developed on a randomly selected subset of 0.01% of the dataset before being preregistered. After preregistration, 36 additional reported test statistics were excluded using non-preregistered rules (3 which reported an r of 1, leading to an infinite Cohen’s z, and 33 test which were parsed by statcheck as reporting impossible test statistics). All random effects meta-regressions were performed with an additional random effect at the lowest level (i.e., at the effect or article level) than was preregistered, as this was thought to be more conceptually appropriate as all tests within a paper cannot be assumed to have estimated the same parameter. The analyses performed on all data (i.e., those which do not attempt exclude peripheral tests) were estimated using the R package nlme in lieu of metafor {Viechtbauer, 2010 #796} as the memory requirements of metafor exceeded those that we had access to (requiring > 160 gbs of RAM). The results should be identical for all practical purposes (e.g., reperforming the two analyses which looked at just the first reported effect or the largest reported effect in nlme lead to parameter estimates that differed by less than .000001 from that produced using metafor, and the standard errors of each estimated parameter differed by less than 0.005). Any statistical tests that were not preregistered are identified as such below.

### Limitations

Before turning to the results of this analysis, it should be noted that the sample is limited to articles published in 5 APA journals in a limited time period (1985 - 2013) and the results may not generalize outside of this sample. It is possible for example that publication patterns have changed and studies with larger effect sizes tend to target other publications (or the reverse). Additionally, the method used to collect effect sizes from the literature, using regular expressions to extract statistical tests has its limitations. This method only captures statistical tests results reported in-text (i.e., not in tables) in APA style, and the observed effects could be driven by changes in how likely people are to report statistical tests in APA style (e.g., if people have become more likely to report small or non-significant results in text in full APA style over time).

## Results

### Descriptives

The mean effect size reported across the sample in correlation coefficient terms is 0.335 and the median is 0.29. Overall, the distribution of effect sizes in the meta-analytic subset is close to that seen in the full dataset, although the median and means are slightly higher (at 0.362 and 0.318 respectively). The effect sizes seen when examining only the highest effect size reported in each paper are much higher on average (with a mean correlation coefficient of 0.596 and a median of 0.62). See Table 1 and Figure 1 for a full list of descriptives about, and histograms of, the distribution of effect sizes in each subsample, and Figure 2 for a QQ plot of effect sizes from the Meta-analytic subset against the effects seen in the full sample. It is noteworthy that the mean effect sizes seen across the whole sample are remarkably close to Cohen’s suggested “medium” effect size benchmark value of r = .3, although the upper are the lower and upper quantiles (0.182, 0.447) are, respectively, higher and lower than Cohen’s “Small” and “Medium” effect size benchmarks (.1 and .5), an occurrence that has been noted in other subfields of research {Quintana, 2017 #836}.

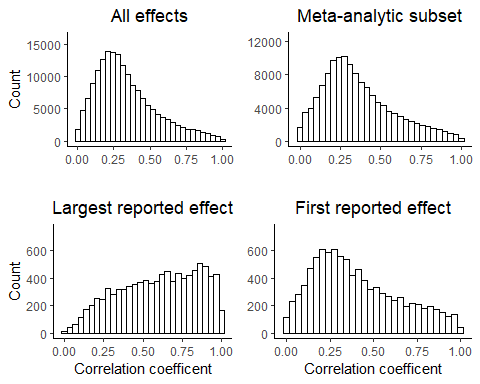


Figure 1. Histograms of reported effect sizes transformed to correlation coefficients, for all results which could be transformed to correlation coefficients, the meta-analytic subset, the the first reported effect size in each paper, and the largest reported effect size in each paper.

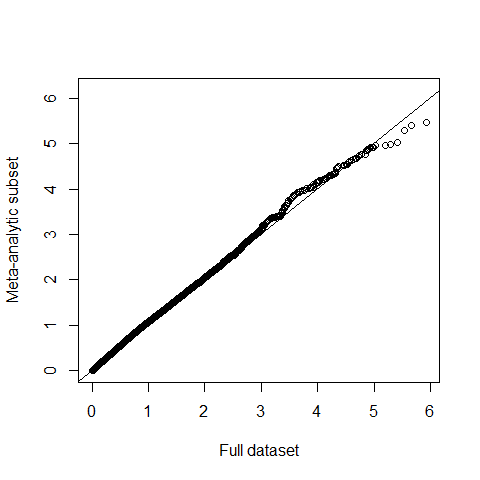


Figure 2. A qq plot of effect sizes in scores from all data plotted against the meta-analytic subset.

Table 1. Descriptives of the reported effect sizes in this sample in and correlation coefficient terms.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subsample | Effect size | n | Mean | sd | Min | 25th percentile | Median | 75th percentile | Max | Range | Skew | Kurtosis |
| All data | Correlation | 170556 | 0.335 | 0.209 | 0.000 | 0.182 | 0.290 | 0.447 | 1.00 | 1.000 | 0.894 | 0.351 |
| All data | Fisher z | 170556 | 0.386 | 0.332 | 0.000 | 0.184 | 0.299 | 0.481 | 6.18 | 6.183 | 3.027 | 18.826 |
| Meta-analytic subset | Correlation | 132086 | 0.362 | 0.218 | 0.000 | 0.203 | 0.318 | 0.489 | 1.00 | 1.000 | 0.747 | -0.010 |
| Meta-analytic subset | Fisher z | 132086 | 0.426 | 0.358 | 0.000 | 0.206 | 0.330 | 0.534 | 6.18 | 6.183 | 2.830 | 16.413 |
| Largest reported effect | Correlation | 9472 | 0.596 | 0.251 | 0.001 | 0.398 | 0.620 | 0.815 | 1.00 | 0.999 | -0.269 | -0.996 |
| Largest reported effect | Fisher z | 9472 | 0.842 | 0.575 | 0.001 | 0.421 | 0.725 | 1.143 | 6.18 | 6.182 | 1.598 | 5.422 |
| First reported effect | Correlation | 9472 | 0.404 | 0.245 | 0.000 | 0.211 | 0.351 | 0.572 | 1.00 | 1.000 | 0.558 | -0.597 |
| First reported effect | Fisher z | 9472 | 0.500 | 0.430 | 0.000 | 0.214 | 0.367 | 0.650 | 4.68 | 4.681 | 2.109 | 7.115 |

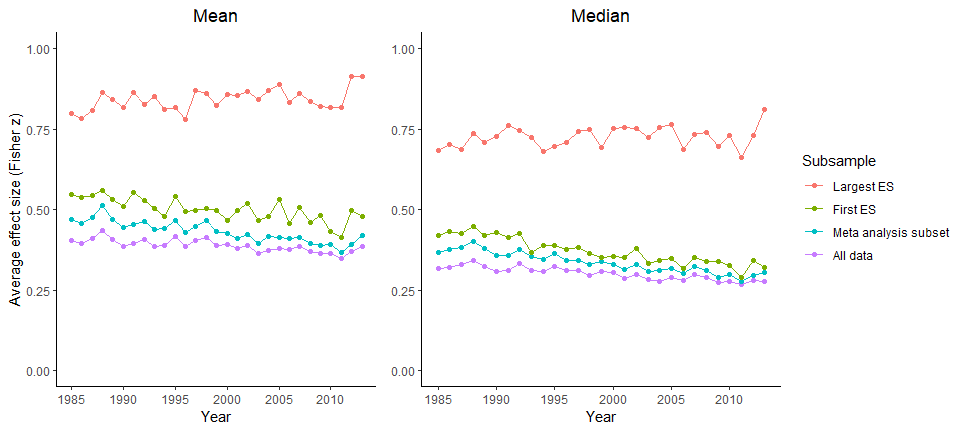


Figure [ESs over time]. Plots of the mean and median effect sizes per year (in Fisher Z transformed correlation coefficients) by the subsamples used in analyses.

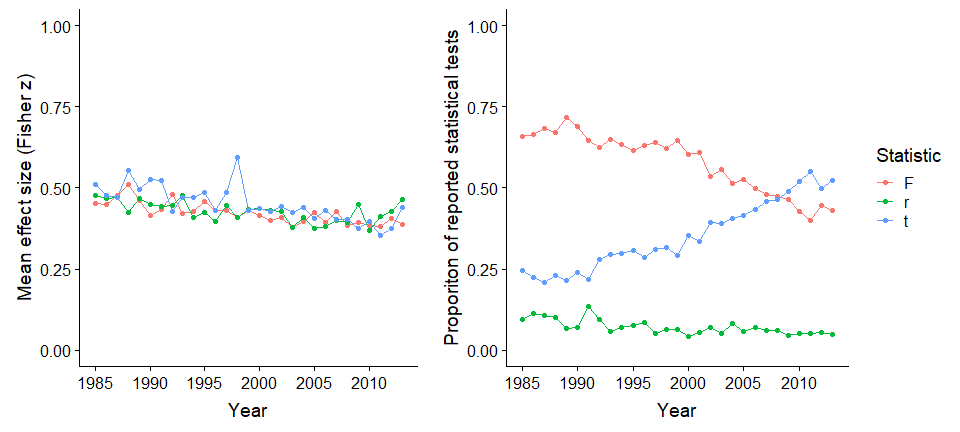


Figure [effectByStat]. A plot of the mean effect sizes per year (left, in Fisher Z transformed correlation coefficients) as transformed from the various effect size measures, and of the proportion of reported statistical tests of each type included in the current analysis (right).

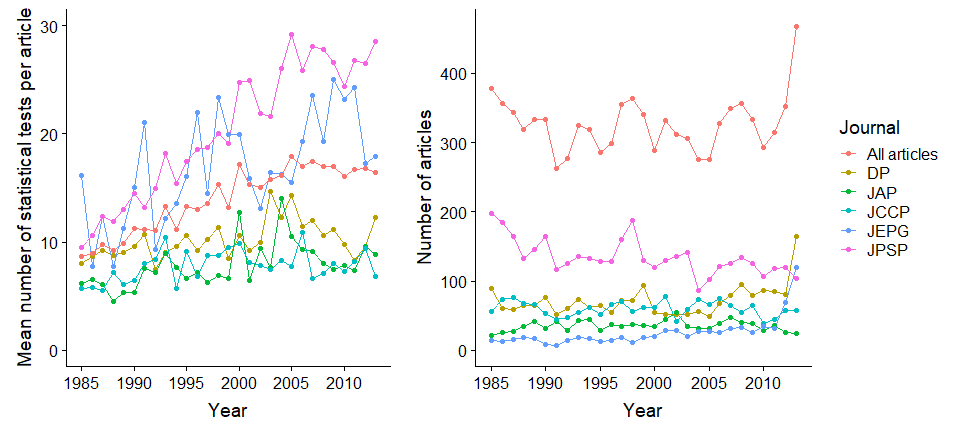


Figure [n tests]. A plot of the mean number of tests reported in each article (left), and the number of articles reported by journal and overall (right).

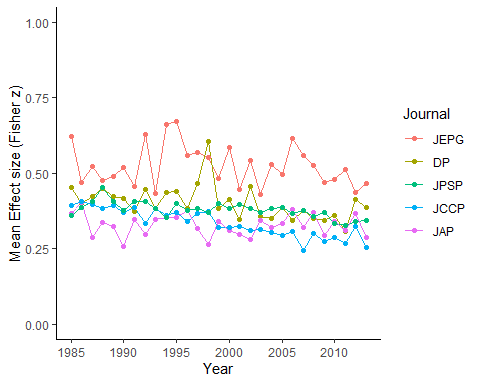


Figure [trends by J]. A plot of the mean effect size reported in each journal by Year.

The number of t, F and correlational statistical tests reported in APA style per article has increased considerably over time in this sample (i.e., of articles that reported at least one), from a mean of 9.578 reported per article in 1985 - 1990, to a mean of 16.59 from 2009 - 2013 (note that these averages only include articles with at least one reported statistical test). See figure [n tests] for a plot of the mean number of tests reported per article over time, and a plot of the number of articles reported in each journal included in this current analysis.

The proportion of t tests in this sample has increased over time, relative to F tests and correlations, see figure [effectByStat]. There are no obvious differences between the average reported effect size in each unit (mean for correlations = 0.427 , mean for F statistics = 0.421, mean for t statistics = 0.434), and the trend over time appears to be consistent among all sources of effect sizes (see figure [effectByStat]).

### Results

A non-preregistered analysis shows that there is a low correlation between year of publication and effect size () of r(170556) =-0.074, p < .001, 95% CI [-0.079, -0.069]. When averaging the effect sizes seen in each article to avoid issues of non-independence of statistical tests within articles and estimating the correlation between year and effect size we find a very weak association between effect size and year of publication, r(10401) = -0.081, p < .001, 95% CI [-0.1, -0.061].

The multilevel meta-regression including all data shows an estimated decrease per year of -0.004, 95% CI [-0.005, -0.004]. Unsurprisingly given that articles likely report multiple statistical tests of different hypotheses in each article there is a large amount of unexplained heterogeneity in effect sizes, QE(132085) = 38658.96, p = < .001, = 97% {Nakagawa, 2012 #1023}. This suggests that 97% of residual variance in effect sizes is due to effect size heterogeneity (i.e., variance in the true effect size differences), while the remaining 3% is attributable to sampling variance. More variance is attributable to the article and effect level than to the journal (with estimated standard deviations for each random effect of = 0.182, 95% CI [0.179, 0.186], = 0.231, 95% CI [0.23, 0.232], compared to = 0.075, 95% CI [0.037, 0.149]), representing a very low interclass correlation (ICC) for the journal of 0.06, with most variance in true effect size differences being accounted for at the article and effect level (with an ICC for the article of 0.36 and an ICC for the effect of 0.58 ).

Table [nice MLME sum datMeta]. Multilevel meta-regression output including all applicable effects. = 132086, = 9472.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.412 | 0.346 | 0.478 | 0.033 | < .001 |  |
| Year | -0.004 | -0.005 | -0.004 | 0.000 | < .001 |  |
|  |  |  |  |  |  | Effect variance = 0.053, n = 132086 |
|  |  |  |  |  |  | Article variance = 0.033, n = 9472 |
|  |  |  |  |  |  | Journal variance = 0.006, n = 5 |
|  |  |  |  |  |  | QE(132085) = 38658.96, p = < .001 |

#### Exploratory analyses

Including only the first reported statistical test in each paper provides similar results, suggesting a small decrease over time, with a -0.003 (95% CI [-0.004, -0.002]) estimated yearly change according to the model including only the first reported effect effects. The estimated value (96.72%) is substantively identical to that of the model including all applicable effects. Looking at the variance partitioning, more variance is attributable to the article level than to the journal level ( = 0.371, 95% CI [r], compared to = 0.082, 95% CI [r]), representing a ICC for the journal of 0.047, with most variance accounted for at the article level (ICC for the article 0.953).

Table [nice MLME sum datMetaFirst]. Multilevel meta-regression output including the first reported effect size in each article, = 9472.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.484 | 0.411 | 0.557 | 0.037 | < .001 |  |
| Year | -0.003 | -0.004 | -0.002 | 0.000 | < .001 |  |
|  |  |  |  |  |  | Journal variance = 0.007, n = 5 |
|  |  |  |  |  |  | Article variance = 0.138, n = 9472 |
|  |  |  |  |  |  | QE(9471) = 190471.31, p < .001 |

Including only the largest effect reported in APA style in each paper leads to a different conclusion, a predicted yearly increase of 0.0015 (95% CI [0.0002, 0.0028]). Again, the estimated value (97.9) is functionally identical to those of the dataset including all data, and more variance is again attributable to the article level than to the journal level ( = 0.509, compared to = 0.194), leading to a low interclass correlation (ICC) for the journal of 0.127, with most variance in accounted for at the article level (with an ICC for the article of 0.873).

Table [nice MLME sum datMetaLargest]. Multilevel meta-regression output including just the largest reported effect size in each article, = 9472.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.815 | 0.645 | 0.986 | 0.087 | < .001 |  |
| Year | 0.002 | 0.000 | 0.003 | 0.001 | 0.02 |  |
|  |  |  |  |  |  | Journal variance = 0.038, n = 5 |
|  |  |  |  |  |  | Article variance = 0.259, n = 9472 |
|  |  |  |  |  |  | QE(9471) = 306395.25, p < .001 |

### Discussion

Overall, there was a small decrease in the mean effect sizes seen across the examined time period; going from a mean reported effect of r = 0.403 between 1985 - 1990, to a mean reported effect size of 0.335 in last five years included in this dataset, 2009 - 2013. The results of the random effects meta-regression accounting for random effects nested within articles and journals shows an estimated yearly decrease in effect sizes of -0.004 (95% CI [-0.005, -0.004]) in units. This corresponds to an estimated correlation coefficient decrease of -0.10 from 1985 to 2013.

There are a number of possible explanations for this decrease. Firstly, it is possible that the large, obvious, effects in psychology have already been discovered, and that this means that psychologists are now turning to study smaller, more subtle effects. If this is the case, and effect sizes under study in the scientific literature have in fact decreased, in combination with the findings from chapter [meta], this suggests that the average statistical power of psychological research has decreased over time.

Another possible explanation is that the change in average effect sizes could be driven by changes in reporting and analysis practices as opposed to changes in the magnitude of the effects under study. Although the proportion of papers that report at least one statistically significant result has barely changed over time (see Figure [at least one significant result]), there has been a slight decrease in the proportion of reported statistical tests that are statistically significant (see Figure [Significant results over time]). If reporting non-significant results has become more common over time, this could lead to a decrease in the average reported effect size as small results that otherwise might not have been represented in the literature may be available.

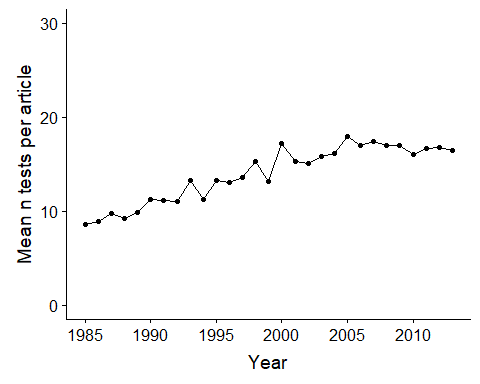


Figure [Significant results over time]. A plot of the mean number of statistical tests reported per article.

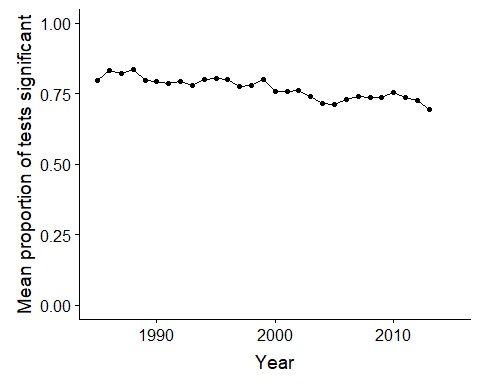


Figure [Significant results over time]. A plot of the mean proportion of reported statistical tests that are significant over time. The proportion of significant tests was calculated at the article level and then averaged for display here.

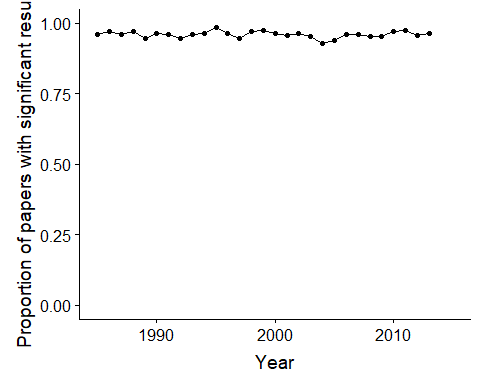


Figure [Significant results over time]. A plot of the number of papers that report at least one significant result over time.

The downward trend in effect sizes is also seen if we only analyze the first reported statistical test in each paper, going from an average correlation of 0.44 between 1985 - 1990 to a mean reported correlation of 0.371 from 2009 to 2013. According to the results of the meta-regression including only the first APA reported result in each paper, there is an estimated yearly decrease of 0.003 (95% CI [-0.004, -0.002]) in units. Over the 28 year time period included in this database, this represents an estimated decrease of -0.07 in correlation coefficient terms.

However, looking only at the largest reported effect in each paper, this trend is no longer apparent. Results from the multilevel meta-regression show an estimated yearly increase of 0.002 (95% CI [0.0002, 0.0028]) in units, or alternatively an estimated increase of r = 0.02 between 1985 and 2013. This result may be driven by the increasing number of statistical tests reported per article (see figure Figure [n tests]). If more analyses are being performed over time (and assuming that the tests performed are at least somewhat independent), selecting the largest reported effect out of each article should show an increased average effect size on the basis of sampling variability alone. Furthermore, the average effect sizes seen when selecting the largest effect in each paper are much larger than have previously been seen in examainations of effect sizes in psychological research (e.g., Bosco et al., 2015; Haase et al., 1982; Paterson et al., 2015; Richard et al., 2003), suggesting that this selection method does not reliably identify the effects of substantive interest. Nonetheless, the results of these exploratory analyses reinforce the fact that the results seen across all reported effect sizes should be interpreted carefully. Future research is necessary before any strong conclusions are drawn about trends in the effects of substantive interest in psychological research.

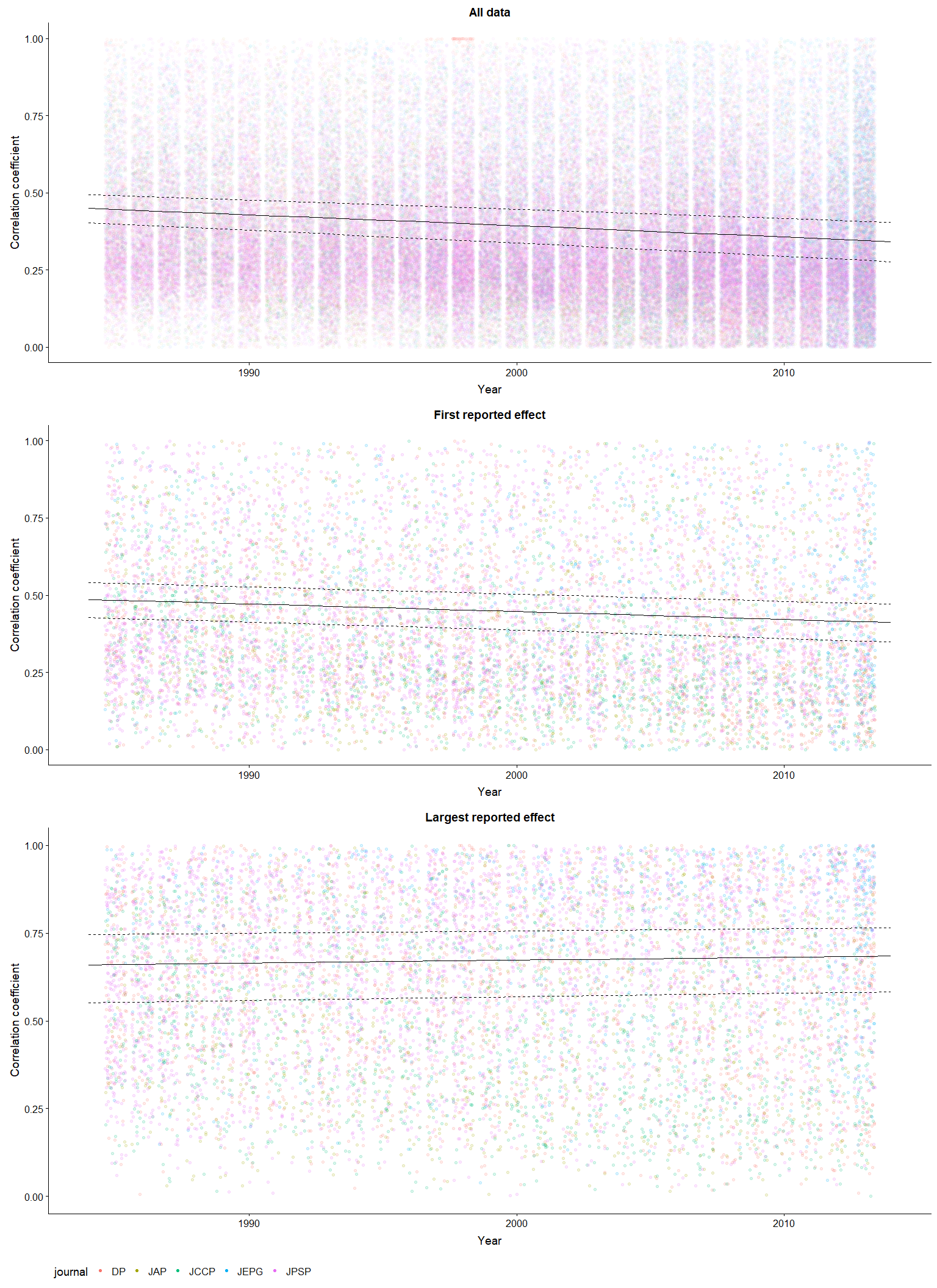


Figure [jitter] A jitter plot of the reported effect sizes in this dataset plotted over time, with an overlaid multilevel meta-regression plot (see Table [nice MLME sum datMeta] for model parameters, the plotted output has been converted to correlation coefficient units).

Overall, this analysis suggests that there has been a substantial decrease in the average sizes of the effects reported in psychology research papers from 1985 to 2013. However, the degree to which this decrease is reflective of a decrease in the size of the focal or main effects under study in psychology remains an open question. Whether these results are representative of focal analyses or not, there is little indication that effect sizes are notably rising over time, a result which supports recent research arguing that the average statistical power of psychology research has remained consistently low across time [meta-analysis], and highlights the need for researchers to consider what effect sizes they expect to see during the planning of their studies in order to avoid unknowingly performing underpowered or imprecise research.

## Supplementary materials

### Supplementary materials 1. Effect size conversions

All statistical tests extracted were transformed into correlation coefficients as follows, using the methods reported in {Open Science Collaboration, 2015 #611}.

t statistics:

Where is the observed t statistic and is the degrees of freedom of the t test.

F statistics:

Where is the observed F statistic and is the degrees of freedom of the numerator and is degrees of freedom of the denominator.

Chi square statistics:

Where is the observed statistic and is the associated degrees of freedom.

All values were then transformed into fisher Z transformed correlation coefficients using:

Standard errors for these statistics when derived from F tests with denominator degrees of freedom of 1, t tests, and correlation coefficients were estimated as:

### Supplementary materials 2. Additional exploratory analyses

##### Ignoring sampling variance and including all data

An alternative approach to analyzing this data is to ignore the sampling variances of each reported effect size and analyse the data using a typical multilevel model with a main effect for year and random effects for article and journal. In this approach, we no longer include random effects at the individual effect-size level, as we could not estimate both residual and effect level variances (both being random effects at the individual effect-size level).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.37982 | 0.31568 | 0.44396 | 0.03273 | < .001 |  |
| Year | -0.00268 | -0.00314 | -0.00222 | 0.00024 | < .001 |  |
|  |  |  |  |  |  | Residual variance = 0.078, n = 132086 |
|  |  |  |  |  |  | Article variance = 0.031, n = 9472 |
|  |  |  |  |  |  | Journal variance = 0.005, n = 5 |
|  |  |  |  |  |  | QE(132085) = 162770.16, p = < .001 |

Using this approach lead to a less extreme year by year difference with the estimated change per year changing by a of 0.00156. Otherwise the model is quite similar. Otherwise, this model shows similar results.

##### Fixed effects for statistic type

It also is possible to estimate fixed effects for each of the included statistical tests (F, r and t tests). Given that these different statistical approaches would be used under different scenarios, this approach seems reasonable. This analysis is otherwise identical to that presented in the main text (i.e., it examines the 132086 effects for which valid standard errors were developed and accounts for the imprecision in each estimated effect size).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.39769 | 0.33175 | 0.46363 | 0.03364 | < .001 |  |
| Year | -0.00447 | -0.00499 | -0.00396 | 0.00026 | < .001 |  |
| r | 0.05388 | 0.04588 | 0.06188 | 0.00408 | < .001 |  |
| t | 0.02719 | 0.02283 | 0.03154 | 0.00222 | < .001 |  |
|  |  |  |  |  |  | Effect variance = 0.053, n = 132086 |
|  |  |  |  |  |  | Article variance = 0.033, n = 9472 |
|  |  |  |  |  |  | Journal variance = 0.006, n = 5 |
|  |  |  |  |  |  | QE(132085) = 38681.93, p = < .001 |

Including fixed effects for statistic type (i.e., a dummy coded variable for r and t tests, leaving F tests as the comparison group) lead to a negligible change in the estimated change over time (with the estimated change per year increasing by -0.00023), but there were small but noticeable effects for effect size type (for t, = 0.02719, 95% CI [0.02283, 0.03154], and for r = 0.05388, 95% CI [0.04588, 0.06188]), with F as the comparison or baseline group (with an intercept of 0.39769, 95% CI [0.33175, 0.46363]))).

##### Allowing slopes to vary by effect size type

Another question we investigated is whether the change in effect sizes over time may differ for the different types of statistical test reported. In order to examine this possibility we included both main fixed effects for statistical test type and interactions between year and test statistic type.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.39767 | 0.33174 | 0.46360 | 0.03364 | < .001 |  |
| Year | -0.00452 | -0.00508 | -0.00396 | 0.00029 | < .001 |  |
| r | 0.05414 | 0.04612 | 0.06215 | 0.00409 | < .001 |  |
| t | 0.02730 | 0.02289 | 0.03171 | 0.00225 | < .001 |  |
| r \* year | 0.00047 | -0.00049 | 0.00142 | 0.00049 | 0.336 |  |
| t \* year | 0.00002 | -0.00053 | 0.00057 | 0.00028 | 0.939 |  |
|  |  |  |  |  |  | Effect variance = 0.053, n = 132086 |
|  |  |  |  |  |  | Article variance = 0.033, n = 9472 |
|  |  |  |  |  |  | Journal variance = 0.006, n = 5 |
|  |  |  |  |  |  | QE(132085) = 38681.5, p = < .001 |

Allowing slopes to vary by statistic type (i.e., including an interaction between dummy coded binaries for r and t statistics) led to negligible estimated effects interaction effects. The interaction effect for correlations, , was 0.00047, 95% CI [-0.00049, 0.00142]. The interaction effect for t statistics, , was 0.00002, 95% CI [-0.00053, 0.00057]. , which now represents the estimated change per year in effect sizes for F statistics, was estimated as -0.00452, 95% CI [-0.00053, 0.00057]. This is only -0.00028 units more extreme than the estimated when including main effects or interaction effects for test statistic types.

##### Allowing slopes to vary by journal

An alternative approach to analyzing this data set is to allow the relationship between effect size and time to vary by journal.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.41282 | 0.34254 | 0.48309 | 0.03586 | < .001 |  |
| Year | -0.00455 | -0.00615 | -0.00296 | 0.00081 | < .001 |  |
|  |  |  |  |  |  | Effect variance = 0.053, n = 132086 |
|  |  |  |  |  |  | Article variance = 0.033, n = 9472 |
|  |  |  |  |  |  | Journal variance = 0.006, n = 5 |
|  |  |  |  |  |  | Slope variance = 0, n = 5 |
|  |  |  |  |  |  | QE(132085) = 38656.75, p = < .001 |

Including random slopes for journal lead to almost no change in the overall estimated change per year ( = -0.00424, a change of just 0.00031), and a very small estimated slope variance of 0.000003, 95% CIs 0.000001, 0.000013. This suggests that the effect size change per year is relatively stable across journals, although the sampling plan used here (only including 5 APA journals), may limit the generalizability of these results outside of this sample.