Effect Sizes Over Time

Felix Singleton Thorn

10 December 2018

Abstract: This article uses a mutlilevel meta-regression framework to estimate the change in average effect size in psychological reserach from 1985 to 2013 using a database of over 130,000 effect size estimates from over 9,000 articles published in APA journals from 1985 to 2013 {Nuijten, 2015 #550}. The results of this analysis suggest that the average effect size reported in psychological reserach has decreased, over time, with the average effect reported in these articles decreasing from 1985 to 2013 by an estiamted -0.004 Fisher Z tranformed correlation coefficent points per year, representing an estimated correlation coefficientdecrease of -0.088. However, methods to isolate the “main effect” studied in each paper suggets that this overall pattern may {hold or not hold}.

## Introduction

An important question in understanding the history of psychological science is whether effect sizes have changed over time. Are we studying smaller effects over time, having already studied the “low hanging fruit” {Baumeister, 2016 #1022}? Or are methodological reforms and the increasing awareness of the importance of measurement in reserach leading to increased effect sizes {citation}? This paper uses an extensive database of over 130,000 effect size estimates from over 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 reserach over this period.

The question of how effect sizes have changed over time has important implications for understanding the results of large scale power surveys of psychological research (e.g., {Cohen, 1962 #487;Rosnow, 1989 #37;Szucs, 2017 #25}), as comparisons of these analyses over time assume that effect sizes have been consistent over time. Most of the efforts to estimate the statistical power of psychological reserach have used Cohen’s effect size benchmarks, values he cautions researchers to be weary of {Cohen, 1988 #562}. According to a recent meta-anlaysis of 46 power surveys of psychological research [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 these values over time. However, in order to know whether the statistical power of psychological research has changed over time, it is necessary to know whether the effect sizes under study in psychological research have changed.

A small number of pervious studies have extracted effect size benchmarks from psychological reserach in domains as varied as social psychology {Richard, 2003 #603} and management psychology {Paterson, 2015 #817; Bosco, 2015 #157}, showing estimates of the average effect size in various subfields which range from an mean correlation coefficient of .21 seen 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 published in high impact journals {Szucs, 2017 #25}. However, we have not identified any studies which have adequate 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}, which 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 look for changes over time across fields of psychological reserach.

### 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 which were reported in APA style from a total of 16695 articles. This paper only uses a subset of these studies, those for which results were avaliable going back to 1985. This analysis therfore includes data from five psychology journals chosen to be representative of the main subdisiplines of psychology reserach; 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 coefficents (henseforth ) following {Open Science Collaboration, 2015 #611} for visulisation and analysis, see supplementary materials 1 for the transformations used. Negative effect sizes (i.e., negative correlations) were set to be positive for ease of analysis and visulisation. Standard errors were estimated as , taking the degrees of freedom for the denominator minus two as n (or 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 standardised formal (and therefore this information was not avaliable in this dataset), Z scores were excluded from analyses (n = 7539). As it was not possible to derrive valid standard errors for F statistics with effect degrees of freedom above 1 (n = 19585) 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 derrived from the entier sample). An additional subset of results were excluded from all analayses and visulisations as they produced standard errors or Z transformed correlation coefficents which could not be estimated or were infinite (n = 19698, 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 of < 6). 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 (). Because of the high memory capacity necessary to perform these analyses, the University of Melbourne’s Spartan high performance computing platform was used to estimate these multilevel models {Meade, 2017 #1020}.

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 all standard errors as above (i.e., for F tests and for tests), a total of n = 170324 effects after excluding all invalid results (i.e., analyses where it was not numerically possible to estimate a standard error or correlation using the above methods). This analysis lead to … (i.e., …!!!). The results below present the results excluding those studies.

### Accounting for peripheral tests

Due to the large sample size included in this anlaysis, there is no feasible way of manually verifying whether the statistical tests included are not, for example, manipulation tests or randomisation tests, and it is almost certain that a large proportion of those tests reported are in fact not tests of the main hypotheses of each paper. This means that any observed change in effect sizes could be driven by changes in reporting practices, for example becoming more or less likely to report manipulation checks or randomisation checks over time. In order to account for this issue, two main methods were used to attempt to identify the focal test of each article. (a) Multilevel mixed effects meta-regression was performed looking just at the first statistical test reported in each paper. (b) Multilevel mixed effects meta-regression was performed using just 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 size analyses was taken.

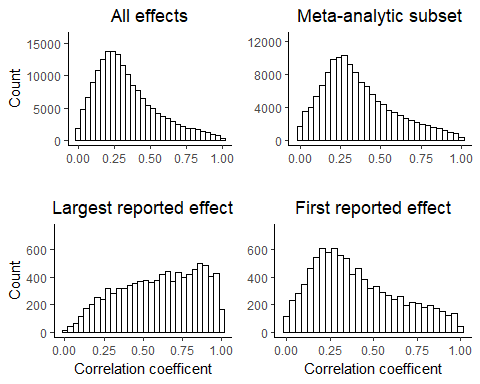
These multilevel-meta-regression includes random effects for individual articles and journals , and includes year of publication of each article as a fixed effect (). These analyses did not include random effects for each statistical test, as each article only provides a single statistical test result. All plots and analyses refering to these subsets only display effects with standard errors, and only those studies with valid standard errors were included in these analyses.

### Deviations from preregistration

All analyses were tested and developed on a subset of 0.01% of the dataset before being pre-registered. 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 values. All random effects meta-regression tests were performed with an additional random effect at the lowest level (i.e., at the effect level or article) 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. Fixed effects change by less than {!!!} if this additional random effect is not included. Any other non-preregistered statistical test are marked as exploratory.

## Results

### Descriptives



The mean effect size reported in this study in correlation coefficent terms is 0.335 and the median is 0.29. Overall, the distribtuion 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 respectivly). The effect sizes seen when examining only the highest effect size reported in each paper are much higher on average (with a mean correlation coefficent of 0.596 and a median of 0.62). See table 1 and and figure 1 for a full list of descriptives about and histrograms of the distribution of effect sizes in each subsample.

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, respectivly, higher and lower than Cohen’s “Small” and “Medium” effect size benchmarks (.1 and .5), an occourance that has been noted in other subfields of reserach also {Quintana, 2017 #836}.

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

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subsample | Effect size | n | Mean | sd | Min | 25th percentile | Median | 75th percentile | Max | Range | Skew | Kurtosis |
| All data | Fisher’s z | 170324 | 0.335 | 0.209 | 0.000 | 0.182 | 0.290 | 0.447 | 1.00 | 1.000 | 0.894 | 0.350 |
| All data | Correlation | 170324 | 0.387 | 0.332 | 0.000 | 0.184 | 0.299 | 0.481 | 6.18 | 6.183 | 3.027 | 18.823 |
| Meta-analytic subset | Fisher’s z | 132086 | 0.362 | 0.218 | 0.000 | 0.203 | 0.318 | 0.489 | 1.00 | 1.000 | 0.747 | -0.010 |
| Meta-analytic subset | Correlation | 132086 | 0.426 | 0.358 | 0.000 | 0.206 | 0.330 | 0.534 | 6.18 | 6.183 | 2.830 | 16.413 |
| Largest reported effect | Fisher’s z | 9472 | 0.596 | 0.251 | 0.001 | 0.398 | 0.620 | 0.815 | 1.00 | 0.999 | -0.269 | -0.996 |
| Largest reported effect | Correlation | 9472 | 0.842 | 0.575 | 0.001 | 0.421 | 0.725 | 1.143 | 6.18 | 6.182 | 1.598 | 5.422 |
| First reported effect | Fisher’s z | 9472 | 0.404 | 0.245 | 0.000 | 0.211 | 0.351 | 0.572 | 1.00 | 1.000 | 0.558 | -0.597 |
| First reported effect | Correlation | 9472 | 0.500 | 0.430 | 0.000 | 0.214 | 0.367 | 0.650 | 4.68 | 4.681 | 2.109 | 7.115 |

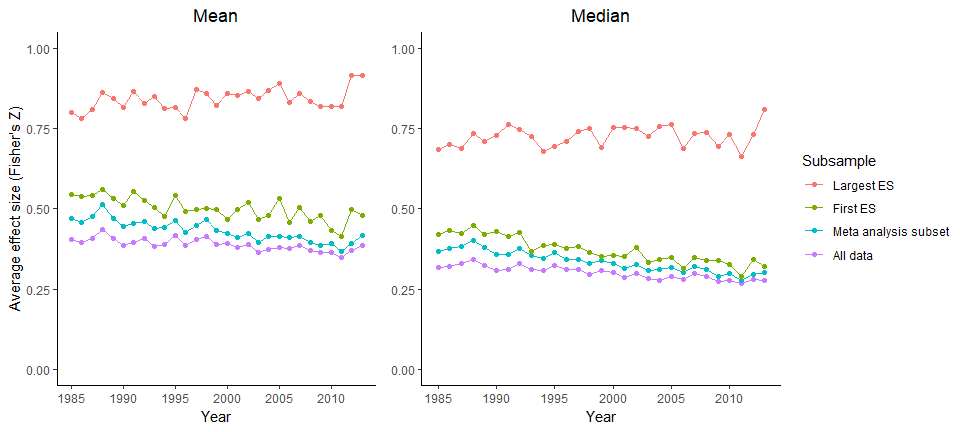


Figure [ESs over time]. Plots of the mean and median effect sizes per year (in Fisher Z transformed correlation coefficents) by the subsamples used in analyses below.

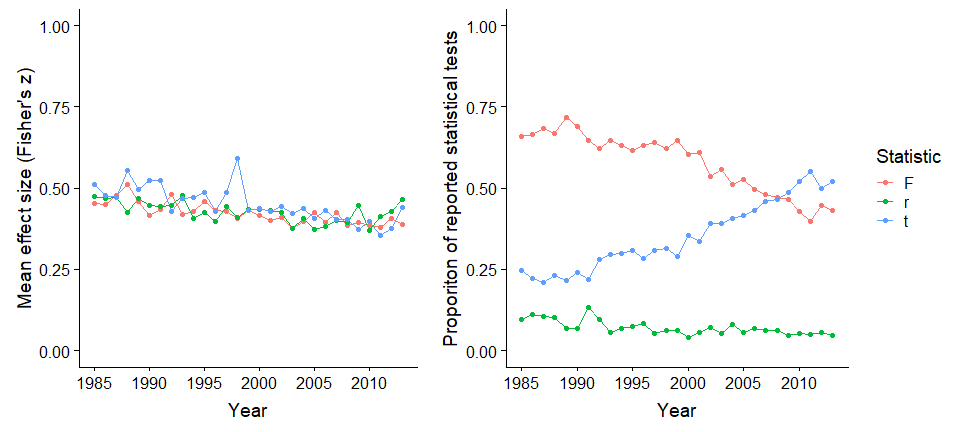


Figure [effectByStat]. A plot of the mean effect sizes per year (left, in Fisher Z transformed correlation coefficents) as transformed from the various effect size measures, and of the proportion of reported statistical tests tests of each type included in the current analysis (right).

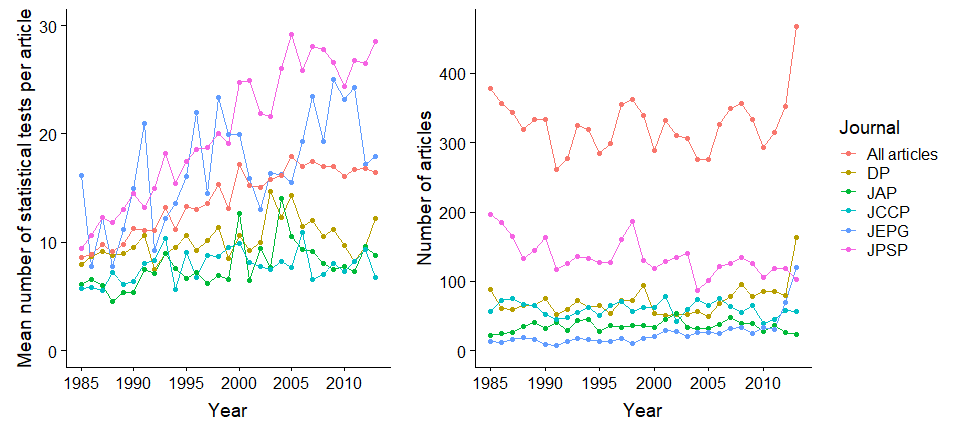


Figure [n tests]. A plot of the mean number of tests reported in each article (left), and the number of articles reported by journal and overall (right).

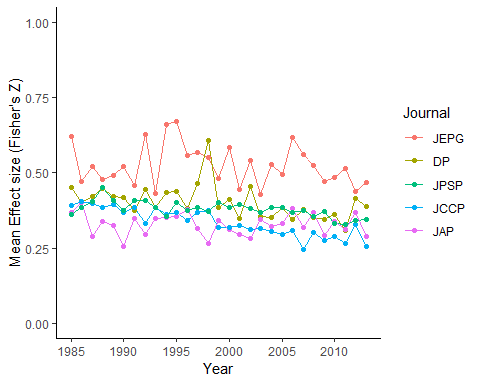


Figure [trends by J]. A plot of the mean effect size reported in each journal by Year.

The 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]).

The number of t, F and correlational statistical tests reported in APA style per article has increased considerably over time in this sample, from a mean of 9.578 reported per article included in 1985 - 1990, to a mean of 16.59 in the final 5 years included in this database (from 2009 - 2013). 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.

### Analysis Results

There is a low correlation between year of publication and effect size () of r(170324) =-0.039, p < .001, 95% CI [-0.044, -0.034, or when averaging within articles r(10399]) = -0.08, p < .001, 95% CI [-0.099, -0.061].

The multilevel meta-regression including estimates a decrease per year of -0.004, 95% CI [-0.006, -0.003]. In correlation coefficent terms this is a one year estimated change of -0.004 95% CI [-0.006, -0.003]. This means that there is an estimated correlation coefficent coefficient decrease of -0.088 in the estimated average effect size from 1985 to 2010, over a 25 year period. Unsurprisingly, there is a large amount of unexplained effect size hetrogeneity QE(2999) = 26448.36, p < .001, and an of -99 {Nakagawa, 2012 #1023}, suggesting that -99% of variance in effect sizes is due to effect size heterogeneity (i.e., variance in the true effect size differences), while the remaining 199% is attributable to sampling variance. More variance is attributable to the article and effect level than to the project ( = 0.017, ( = 0.056, compared to = 0.004), representing a very low intraclass correlation (ICC) for the journal of 0.05, and a low ICC for the article of 0.223.

Table [nice MLME sum datMeta].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| 0.393 | 0.337 | 0.449 | 0.029 | < .001 | NA |
| -0.004 | -0.006 | -0.003 | 0.001 | < .001 | NA |
| NA | NA | NA | NA | NA | Journal variance = 0.004, n = 5 |
| NA | NA | NA | NA | NA | Article variance = 0.017, n = 2264 |
| NA | NA | NA | NA | NA | Effect variance = 0.056, n = 3000 |
| NA | NA | NA | NA | NA | QE(2999) = 26448.36, p < .001 |

DIFFS WITH ALL DATA INCLUDING SEs which are a little shifty: 0.034, -0.002 Table [nice MLME sum datAll].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| 0.359 | 0.328 | 0.39 | 0.016 | < .001 | NA |
| -0.002 | -0.003 | 0.00 | 0.001 | 0.0275321968771344 | NA |
| NA | NA | NA | NA | NA | Journal variance = 0.001, n = 5 |
| NA | NA | NA | NA | NA | Article variance = 0.01, n = 1688 |
| NA | NA | NA | NA | NA | Effect variance = 0.051, n = 2000 |
| NA | NA | NA | NA | NA | QE(1999) = 17379.94, p < .001 |

Including only the first reported statistical test in each paper provides similar results, suggesting a small but noticable decrease over time, with a -0.003 (95% CI [-0.004, -0.002]) estimated yearly change acorrding to the model including only the first reported effect effects. This is equivalent to a correlation coefficent change of -0.003, 95% CI [-0.004, -0.002], or a change in estimated average correlation coefficents between 1985 and 2010 of -0.062. The estimated value (96.721) is functionally identical to those of the dataset including all data, and looking at the variance partitioning more variance is again attributable to the article level than to the journal level ( = 0.138, compared to = 0.007), representing a low intraclass correlation (ICC) for the journal of 0.047.

Table [nice MLME sum datMetaFirst].

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

Including only the largest effect reported in APA style in each paper leads to a different story, a predicted yearly increase of 0.002 (95% CI [0, 0.003]), equivalent to a correlation coefficent change of 0.002, 95% CI [0, 0.003], or an increase in estimated mean correlation coefficients between 1985 and 2010 of r = 0.021. Again, the estimated value (97.905) is functionally identical to those of the dataset including all data. Again, more variance is again attributable to the article level than to the journal level ( = 0.259, compared to = 0.038) leading to a low intraclass correlation (ICC) for the journal of 0.127.

Table [nice MLME sum datMetaLargest].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| 0.815 | 0.645 | 0.986 | 0.087 | < .001 | NA |
| 0.002 | 0.000 | 0.003 | 0.001 | 0.0215906483375469 | NA |
| NA | NA | NA | NA | NA | Journal variance = 0.038, n = 5 |
| NA | NA | NA | NA | NA | Article variance = 0.259, n = 9472 |
| NA | NA | NA | NA | NA | QE(9471) = 306395.25, p < .001 |

### Discussion

Overall, there was a small decrease in the mean effect sizes seen across the time period from a mean reported effect of r = 0.356 between 1985 - 1990, to a mean reported effect size of 0.318 in last five years included in this dataset, from 2009 to 2013. The formal modeling of this dataset, accounting for random effects nested within articles and journals supports this idea, showing an estiamted yearly decrease in effect sizes of 0.004 (95% CI [-0.006, -0.003]) in units (almost identical to correlation coefficent units when below .4). Looking just at the first reported statistical test in each article as a proxy for the main result of each paper, there is also a decrease in the average effect sizes reported over time. There was an average correlation of 0.44 between 1985 - 1990, to a mean reported correlation of 0.371 from 2009 to 2013, and an estimated decrease per year according to the above multilevel meta-regression results of 0.003 (95% CI [-0.004, -0.002]) in units. The estimated amount of change year-to-year is not large by any means, but may represent a meaningful decrease in the types of effects that are being studied in psychological research, especially when considering the relativly small timespan included in this analysis.

However, looking only at the largest reported effect in each paper, this trend is no longer apparent. There was an average correlation of 0.598 between 1985 - 1990, compared to a mean of 0.593 2009 - 2013, a slight increase in the size of the largest reported effect in each paper. Results from multilevel meta-regression show an estiamted yearly increase of 0.002 (95% CI [0, 0.003]) in units. There are several explanations for this result. This effect could be driven in part by a larger number of statistical tests being reported per article (see figure Figure [n tests]). Assuming that all performed tests are reported or at least that the largest observed test result is reported, if more analyses are being performed over time (and assuming that the tests performed are 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. This effect does not appear to be driven by a subset of statistical tests (see Figure [effectbystatLargest]).

Other possible reasons ???.

In any case, the estimated change over time is so small as to be practically dismissible, with an estimated change from 1985 to 2010 of just r = 0.021, an effect small enough to be accounted for by changes in reporting or analysis practices.

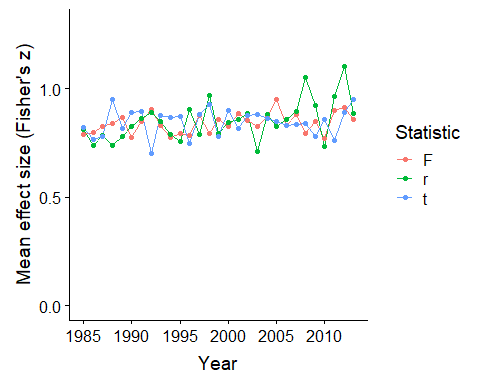


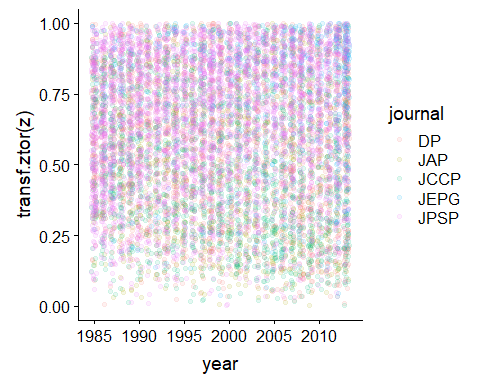
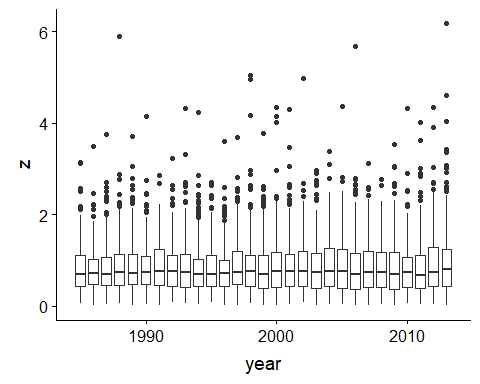
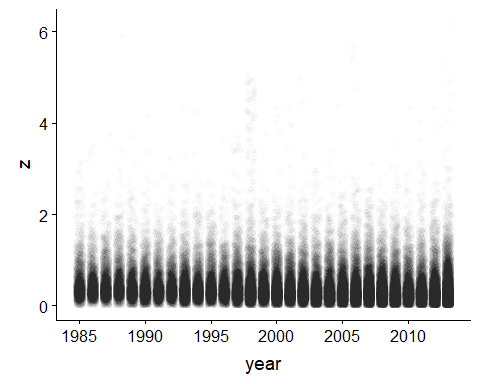
Figure [effectbystatLargest]

Limitaitons

It should be noted that the sample is limited to articles publihsed in 5 APA journals in a limited time period (1985 - 2013). How broadly these findings generalise outside of this population is an open question. It is possilble that recent changes in the design, performance and reporting of experiments (pratices such as preregistration of analyses and and increasing focus on measurement issues) may influence the effect sizes reported in the psychological literature {Nelson, 2018 #750;Nosek, 2018 #887}. Secondly, the method used to collect effect sizes from the literature, using regular expressions to extract statistical tests reported in text in APA style has its limitiations. This method only captures statistical tests results reported in-text (i.e., not in tables) in APA style. If people’s reporting practices are influenced by the size of the effects, this could lead to biased estiamtes of the effect sizes.

It is also worth noting that in ANOVA designs or multiple regression where there is more than one factor or covariate included, the effect size conversion necessarily leads to the exclusion of any non-focal variables’ variance. This means that a study which includes a variable as a covariate will lead to a larger observed effect size than a study which does not include the covariate, although the relationship between the the focal variable remains constant {Olejnik, 2003 #933}. When the extracted effect sizes are based on F statistics, changes in the observed relationships over time could be caused by changing habits in the use of variables in regression or ANOVA designs. However, the trend amoung different statistical tests results appears to be consistent across statistical tests (see Figure [effectByStat]), and for many purposes (e.g., power analysis) the effect size of interest is accurately represented by this value.

Finally, given the effect of publication bias in psychological literature, it is unclear whether the results reported in journal articles are representative of the results that should be expected in planning an experiment {Fanelli, 2010 #222}{Rosenthal, 1979 #490}, and extreme caution is advised before they are used for this purpose. A recent re-analysis of all of the large scale replication studies (such as {Open Science Collaboration, 2015 #611}) suggests that the amount of effect size inflation seen in non-clinical behavioural science is approximately 19%, with a 95% highest probability density interval of 11% to 28% [cite publication bias paper].



### Conclusion

Looking at the

## Appendix

### Conversion

All statistical tests extracted were transformed into correlation coefficents following the methods reported in {Open Science Collaboration, 2015 #611}. t statistics were converted using:

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

F statistics were converted using:

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

And chi square statistics as

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

And standard errors for these statistics when derrived from F tests with denominator degrees of freedom of 1, t tests, and correlation coefficents were estimated as: