To do:

Extract all author + paper information, process and put into a table for each article

Extract all test statistic + effect size information, process and put into table for each article

Decide on tables to go in database – currently:

* Metadata
  + PMCID [character] (Pub Med Central ID, used to ID article in each table)
  + doi [character] (doi)
  + journalID [character] (journal name)
  + journalIDAbrev [character] (abbreviated journal ID)
  + title [character] (title of article)
  + issue [numeric] (issue number)
  + volume [numeric] (volume number)
  + pPub [date] (print publication date)
  + ePub [date] (electronic publication date)
  + abstract [character] (full text of abstract)
  + call [character] (URL of call to XML of article)
* Authors
  + PMCID [character]
  + firstName [Character] (author first name)
  + lastName [Character] (author last name)
* Keywords
  + PMCID [character]
  + keywords [character] (keywords)
* Statistics
  + PMCID [character]
  + section [character (one of: abstract, intro, methods, discussion, results, unlabelled)] (the section from which the statistic was separated into)
  + statistic [character (one of: t, F, r, chi, d, eta, HR, OR)] (the statistic extracted from the paper, t statistic, F statistic, r (correlation coefficient), chi squared (chi), Cohen’s d, eta (eta squared, eta, partial eta, omega squared, epsilon squared), Hazard Ratio (HR), Odds Ratio (OR)
  + cleaned [character] (the extraction section from the paper, all whitespace removed)
  + reported [character] (the extraction section from the paper, copied exactly as reported)
  + context [character] (100 words around the extracted statistic)
  + value [numeric] (test statistic or effect size value)
  + df1 [numeric] (effect degrees of freedom for F tests, NA otherwise)
  + df2 [numeric] (residual / error degrees of freedom)
  + p [character (may include “<” or “>” if reported, equal signs removed)] (p value for each statistical test)

All are NAs if not applicable or not reported

Create SQL database

Training set:

50 articles – extract all test statistics

Validation set:

50 articles – extract all test statistics

ALSO write another paper about the trend in effect size reporting – i.e., descriptive and over time.

Also write another paper about the trend in effect sizes reported in the literature over time. Are they in fact going down.

Maybe run them all through statcheck too? Or at least the APA formatted ones.

One of the most useful tools that facilitates succinct and understandable data summaries, comparison of results across studies and future power analysis are standardised effect sizes. Although statistical significance tests are often presented as the primary tests of a scientific theory, equally if not more important is the degree of difference or strength of effect that was observed. Often, the most efficient way of expressing the size of the effect will be to use a standardised effect size (e.g., Cohen’s d, partial , R2). These effect size benchmarks have become extremely tightly engrained in much of the discussion around effect sizes, especially in psychology. Many of the most commonly used benchmarks to understand and planning effect sizes in psychology are based on Cohen’s (1988) benchmarks. This is true despite the fact that Cohen never suggested using his benchmarks as anything but a last resort in power analysis or effect size interpretation.

E.g.; “When the above has not provided sufficient guidance, the reader has an additional recourse. For each statistical test's ES index, the author proposes, as a convention, ES values to serve as operational definitions of the qualitative adjectives" small,"" medium," and" large." This is an operation fraught with many dangers … they run a risk of being misunderstood.”, Cohen, 1988, p.12.

And despite the fact that he selected these benchmarks in a somewhat ad hoc manner e.g., “‘Small’ effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task, yet not so large as to make them fairly perceptible to the naked observational eye… In contrast, large effects must not be defined as so large that their quest by statistical methods is wholly a labor of supererogation.” Cohen, 1988, p.13.

This project uses a big data and text mining approach to extract effect sizes reported in standard formats in over in over two million open access articles in the PubMedCentral open access subset, and presents an online tool for accessing, visualising and downloading this dataset. This does not provide a method for estimating the expected outcome of a particular area of psychological research, just of visualising the reported effects in an area of psychology research. It can be used as heuristic guide to understanding effect sizes, but for specific estimates of the expected effect sizes from particular experiments, researchers should look to meta-analyses or other large scale replication curation projects like curatescience.org (LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018).

**Adjusting for publication bias:**

One of the main risks in developing projects like this is the issue of effect size exaggeration due to reporting and publication bias (Ferguson & Brannick, 2012). In order to address this issue, several methods have been implemented to adjust for publication bias.

For effect sizes estimated from test statistics directly …

(see Anderson, Kelley, & Maxwell, 2017; McShane & Böckenholt, 2016; Perugini, Gallucci, & Costantini, 2014; and Taylor & Muller, 1996)

THINK ABOUT ESTIMATING EFFECT SIZES FROM THE TRUNCATED distributions following Anderson & Maxwell

**Selection of journals**

The titles of the 3239 Journals included in the open access subset were hand searched, and all journals taking the topic of the behavioural science were included. This search used a broad definition of behavioural science, including education, cognitive, non-chemical-neuroscience, alcohol and substance abuse, speech pathology and psychiatry journals. However, this search did not include Alzheimer’s and dementia / gentrology journals. The ‘about’ page was checked for all journals where the research domain was not identifiable from the title (e.g., “ESMO Open journals”, “The FASEB Journal”, etc.). After this, journals were excluded if none of the included articles had their full text available. This left x articles, nested within x journals.

See Addressing the “Replication Crisis”: Using Original Studies to Design Replication Studies with Appropriate Statistical Power Samantha F. Anderson & Scott E. Maxwell

Taylor and Muller (1996) suggest using the lower 95% CI of the non-centrality parameter

Anderson and colleagues (2017) method for effect size adjustment estimates the expected effect size by producing a maximum likelihood estimate of the effect size of published given the observed effect size under the assumption that the distribution of reported F statistics is truncated at a given significance level (they suggest .05).

For effect sizes directly extracted, it was not possible to use these methods. In these cases

SUBTASK – calculate average effect size decrease from all of the large reproducibility projects

Effect size information:

Converting F-value from ANOVA to cohen's d in meta-analysis (metafor-package) <https://stat.ethz.ch/pipermail/r-help/2011-July/284933.html>

**Common effect size conversions:**

[**https://www.soph.uab.edu/sites/edu.ssg/files/People/MBeasley/Courses/EffectSizeConversion.pdf**](https://www.soph.uab.edu/sites/edu.ssg/files/People/MBeasley/Courses/EffectSizeConversion.pdf)

Download from PubMedOpen

Plot of the data extraction

# write this chunk in markdown

This set of articles was also examined against the statcheck database

During the development of the program, 25 articles were examined, and the program was altered to detect all statistical tests in these sets where possible (i.e., outside of tables unless the individual statistical test was reported in APA style within the table (e.g., r = .01). When papers reported greater than 15 statistical tests, 15 tests were extracted (chosen non-randomly to represent as many possible variations of statistical test reported in the paper, e.g., “rpb = .03”, “r(23) = .12”, “t(12)=1.2, p = .04” etc) and any alterations required to capture these values were made. As each paper was validated, the results were saved to a test file and for each future alteration it was ensured that the results did not prevent previously captured and validated results from being captured.

These manually extracted values were added to a table, which was used as a test for each paper – i.e., the extraction was changed based on these first 50 papers, and then was confirmed on a further 50 which were also manually extracted.

The 100 papers to be examined were extracted in random order.

Statcheck, which extracts all tests statistics with associated p values was used to assess – [report the sensitivity and specificity of statcheck, and then compare with the current output – we get everything that they get + more]

Use cases:

As a replacement last resort to the benchmarks seen in psychology as a whole

Allows people to quickly gather benchmarks from an area of psychology research instead of relying on Cohen’s benchmarks

As a more subtle tool to provide some idea of the effect sizes seen in particular areas of psychology

As an educational tool, allowing quick visualisation and exploration of the reported effects in different areas of psychological research

As a side effect of the relatively regular way in which these effect sizes are reported, it is possible for a text-search algorithm to be built to collect and dis

it is possible to perform relatively simple operations to extract them from previously written scientific papers, such as

Extracted values:

Partial eta squared

partial eta squared is estimated from F statistic and dfs as

It is possible to exclude or include the values as desired.

Include an “exclude all values greater with a probability of occurring < .001 given that the mean and SD is true and a normal SD” as an anomaly detection method?

ADDITIONALLY EXTRACT:

Sample sizes, of course, and aspects of the study designs, populations sampled, and researcher identities and affiliations. You’re going to have some awkward issues regarding dependencies in the data, e.g., multiple effect sizes within one study, multiple studies from the same lab.

For a citation about the proportion of studies which report an effect size: (Fritz, Morris, & Richler, 2012)

Limitations:

It should be noted, this project is limited in several ways. One, although the number of articles included in this analysis is large (over 2,000,000 articles), they cannot be said to comprehensively cover all of psychology and biomedicine.

It is possible that open access journals are in some way systematically different from the population of all journals.

The way that the text search function works means that there are definitely errors in this database. Although no false positives were identified in the validation stage of this project, it is possible that some people use one of the same sybols to represent another feature of their dataset, in which case it would appear in this database as a reported effect size. Secondly, this search mechanism does not pick up on

Previous efforts along the same lines:

In order to provide an overview of the effects reported in varied fields, serval previous efforts have attempted to extract and report on the effects that have been seen in various fields of psychology research.

Previous attempts to extract effect sizes reported in the scientific literature (e.g., Haase, Waechter, & Solomon, 1982; Quintana, 2016) have focused on particular sub-fields of psychology research and have not presented their data in a way that can easily be accessed and explored readers. Two previous studies have taken a text-analysis or big data approach to answering this question. The first, Szucs and Ioannidis (2017), focus on articles published in 18 journals frequently cited in cognitive neuroscience and psychology published between Jan 2011 to Aug 2014, not examining as large or broad a population.

Second, Bosco, Singh, Aguinis, Field, and Pierce (2015) examine the first correlation table presented in every article published in the Journal of Applied Psychology or Personnel Psychology from 1980 to 2010. Bosco et al., (2015) include the relation types (e.g., “attitudes-performance” or “behaviour-intentions”) and make their data publicly available. Bosco, Uggerslev, and Steel (2017) have continued this approach with the METABus project. The METABus project analyses the first correlation table reported in articles using a combination of text-extraction and a manual checks by a team of over 25 research assistants. Although their approach, which includes a manual check of every data-point, leads to higher-quality data than the fully automated approach of the current project, their project only assesses relations of particular type (correlations) published in a limited set of journals covering IO and management psychology.

Things to include in application:  
Search by: Author, journal, date range,

Keyword, **Optional stemmed keyword search**

Data organisation

Data is exported in a tidy data format as two files, one file including the PubMedCentral ID, DOI, Journal, Volume and Issue numbers, Publication Date, Article Title, and Authors, and the other file including the effect sizes of each type. The two files are keyed on the PMCID columns.

Table [Export file 1]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| PMCID | DOI | Journal | Issue | Volume | Publication Date | Article Title | Authors | Keywords |
| 4547492 | 10.1016/j.jbtep.2015.05.001 | Journal of Behavior Therapy and Experimental Psychiatry | NA | 49 | 2015.12.01 | Investigating the efficacy of attention bias modification in reducing high spider fear: The role of individual differences in initial bias | Elaine Fox; Konstantina Zougkou;  Chris Ashwin; Shanna Cahill | Spider phobia, Spider fear,  Attentional bias, Cognitive bias modification, Attentional training , Threat detection |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

Table [export file 2]

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PMCID | Effect size type | Effect size | Bias-adjusted ES | Location |
| 4547492 | d | .04 | .01 | results |
| 4547492 | d | 7.5 | 5.5 | results |
| 4547492 | d | 3.27 | 3 | results |
| 4547492 | r | .31 | .2 | discussion |
| 4547492 | r | .06 | .04 | discussion |
| 5794850 | r | .66 | .5 | results |

Possible other sources for effect sizes:

http://olabout.wiley.com/WileyCDA/Section/id-829772.html

References

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science, 28*, 1547-1562. doi:10.1177/0956797617723724

Bosco, F. A., Singh, K., Aguinis, H., Field, J. G., & Pierce, C. A. (2015). Correlational Effect Size Benchmarks. *Journal of Applied Psychology, 100*, 431-449. doi:10.1037/a0038047

Bosco, F. A., Uggerslev, K. L., & Steel, P. (2017). MetaBUS as a vehicle for facilitating meta-analysis. *Human Resource Management Review, 27*, 237-254. doi:<https://doi.org/10.1016/j.hrmr.2016.09.013>

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

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychol Methods, 17*, 120-128. doi:10.1037/a0024445

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*, 2-18. doi:10.1037/a0024338

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29*, 58-65. doi:10.1037/0022-0167.29.1.58

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances In Methods and Practices in Psychological Science*, 2515245918787489. doi:10.1177/2515245918787489

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21*, 47-60. doi:10.1037/met0000036

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*, 319-332. doi:10.1177/1745691614528519

Quintana, D. S. (2016). Statistical considerations for reporting and planning heart rate variability case-control studies. *Psychophysiology, 54*, 344-349. doi:10.1111/psyp.12798

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*, e2000797. doi:10.1371/journal.pbio.2000797

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics - Theory and Methods, 25*, 1595-1610. doi:10.1080/03610929608831787