# Effect Sizes and Confidence Intervals Guide

Matthew B. Jané1, Qinyu Xiao2, Siu Kit Yeung3, Daniel J. Dunleavy4, Lukas Röseler5, Mahmoud Elsherif6, Denis Cousineau7, and Gilad Feldman8

1Department of Psychological Sciences, University of Connecticut

2Department of Occupational, Economic, and Social Psychology, University of Vienna

3Department of Psychology, University of Hong Kong

4College of Social Work, Florida State

5University of Bamberg

6University of Birmingham

7Université d’Ottawa

8University of Hong Kong, Department of Psychology

# Author Note

Matthew B. Jané https://orcid.org/0000-0002-3121-7769

Qinyu Xiao https://orcid.org/0000-0002-9824-9247

Siu Kit Yeung https://orcid.org/0000-0002-5835-0981

Daniel J. Dunleavy https://orcid.org/0000-0002-3597-7714

Lukas Röseler https://orcid.org/0000-0002-6446-1901

Mahmoud Elsherif https://orcid.org/0000-0002-0540-3998

Denis Cousineau https://orcid.org/0000-0001-5908-0402

Gilad Feldman https://orcid.org/0000-0003-2812-6599

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article. The author(s) received no financial support for the research and/or authorship of this article. Thank you to Bo Ley Cheng, Katy Tam, and Kristy for their contributions to this project

Correspondence concerning this article should be addressed to Gilad Feldman, University of Hong Kong, Department of Psychology, Email: gfeldman@hku.hk

# Abstract

This effect sizes and confidence intervals collaborative guide aims to provide students and early-career researchers with hands-on, step-by-step instructions for calculating effect sizes and confidence intervals for common statistical tests used in psychology, social sciences and behavioral sciences, particularly when original data are not available and when reported information is incomplete. It also introduces general background information on effect sizes and confidence intervals, as well as useful R packages for their calculation. Many of the methods and procedures described in this Guide are based on R or R-based Shiny Apps developed by the science community. We were motivated to focus on R as we aim to maximize the reproducibility of our research outcomes and encourage the most reproducible study planning and data analysis workflow, though we also document other methods whenever possible for the reference of our readers. We regularly update this open educational resource, as packages are updated frequently and new packages are developed from time to time in this rapidly changing Open Scholarship era.

*Keywords*: effect size, confidence interval, collaboration, open science, open educational resource

# Effect Sizes and Confidence Intervals Guide

*Note.* This is a constantly updated collaborative guide on effect sizes and confidence intervals. The most up-to-date version of this guide is hosted as a Google Doc at this link: <https://mgto.org/effectsizeguide>. A similar guide on power analysis can be found at: <https://mgto.org/poweranalysisguide>. This guide is shared under the [Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)](https://creativecommons.org/licenses/by-nc-sa/4.0/) license.

## Guidelines for contribution

All are encouraged to contribute to this Guide. Please note that this Guide is in continuous development such that it will remain a work in progress for an indefinite period of time. This is intended because we hope the Guide to always reflect the state of the art on the topics of effect sizes and confidence intervals.

### Notes

* Please use the headings and style as set forth in this document. You can use keyboard shortcuts such as Ctrl + Alt + 1/2/3. The normal text is in Times New Roman font, font size 11. The codes are formatted using the Code Blocks add-on of Google Docs, github theme, font size 8.
* Use the Suggesting mode rather than the Editing mode. Suggesting is now the default mode for this document. Therefore, please do not hesitate to correct mistakes or modify the contents directly.
* Add a comment to the document if you find anything missing or improper, or if you feel that things are better organized in a different way. We appreciate your suggestions. If you have any questions, please also add a comment. We will reply and seek to clarify in the document body.
* Please make proper citations (in APA 7th format) and provide relevant links when you refer to any source that is not your own.

### Credit and authorship

If you believe you have made sufficient contribution that qualifies you as an author, and you would like to be listed as an author of this Guide, please do not hesitate and list your name and contact information below. The administrators (Q.-Y. X., S. K. Y., and G. F.) of this Guide will verify your contribution and add you to the author list. We welcome comments from any person, regardless of whether they want to be an author. You are also welcome to request content to be added to this Guide (please see the Things to add to the guide section in the end).

The authorship order is such that Q.-Y. X. and S. K. Y. will be the first two authors and G. F. will be the last and the corresponding author. All other contributors will be listed alphabetically in the middle and are all considered joint third authors. Contributors are by default given investigation, writing - original draft, and writing - review & editing CRediT authorship roles. It is possible to take on more roles if contributors prefer. Any change in this authorship order rule will have to be approved by all who are already listed as an author.

## Evaluating and Interpreting Confidence Intervals

Effect sizes quantify the magnitude of effects (i.e., strength of a relationship, size of a difference), which are the outcomes of our empirical research. Effect sizes are by no means a new concept. However, reporting them remained largely optional for many years, and only until recently does it become a community standard: scientists now see reporting effect sizes (in addition to the traditional statistical significance) as a must and journals also start to require such reporting. Notably, in 2001 and 2010, The Publication Manual of the American Psychological Association 5th and 6th editions emphasized that it is “almost always necessary”[[1]](#footnote-46) to report effect sizes (Association, 2010, p. 34; see Fritz et al., 2012, which provides a comprehensive summary on history and importance of effect size reporting).

Effects sizes can be grouped in broad categories as (1) raw effect sizes, and (2) standardized effect sizes. The raw effect sizes are summary of the results that are expressed in the same units as the raw data. For example, when kilograms are measured, a raw effect size reports a measure in kilogram. Consider the effect of a diet on a treatment group; a control group receives no diet. The change in weight can be expressed as the mean difference between the group. This measure is also in kg and so is a raw effect size. Standardized effect sizes are expressed on a standardized scale which has no longer any unit but which have a universal interpretation. A z score is an example of a standardized measure. This document is concerned exclusively on standardized effect sizes.

## Benchmarks

What makes an effect size “large” or “small” is completely dependent on the context of the study in question. However, it can be useful to have some loose criterion in order to guide researchers in effectively communicating effect size estimates. Jacob Cohen (1988), the pioneer of estimation statistics, suggested many conventional benchmarks (i.e., how we refer to an effect size other than using a number) that we currently use. However, Cohen (1988) noted that labels such as “small”, “medium”, and “large” are relative, and in referring to the size of an effect, the discipline, the context of research, as well as the research method and goals, should take precedence over benchmarks any time it’s possible. There are general differences in effect sizes across different disciplines, and within each discipline, effect sizes differ depending on study designs and research methods (Schäfer & Schwarz, 2019) and goals; as Glass et al. (1981) explains:

Depending on what benefits can be achieved at what cost, an effect size of 2.0 might be “poor” and one of .1 might be “good.”

Therefore, it is crucial to recognize that benchmarks are only general guidelines, and importantly, out of context. They also are tend to attract controversy (Glass et al., 1981; Harrell, 2020; Kelley & Preacher, 2012). Note that empirical benchmarks have been suggested by researchers. For social psychology, these alternative benchmarks obtained through meta-analyzing the literature (for example, [this](https://doi.org/10.1037/1089-2680.7.4.331) and [this](https://doi.org/10.1016/j.paid.2016.06.069); see [this Twitter thread](https://twitter.com/cjsotomatic/status/1144701540839698432) for a summary) are typically smaller than what Cohen put forward. Please refer to the table below:

| Effect Size | Reference | Small | Medium | Large |
| --- | --- | --- | --- | --- |
| *Mean Differences* |  |  |  |  |
| Cohen’s or Hedges’ | Cohen (1988)[[2]](#footnote-51) | 0.20 | 0.50 | 0.80 |
|  |  | 0.18 | 0.37 | 0.60 |
|  | Lovakov and Agadullina (2021)[[3]](#footnote-52) | 0.15 | 0.36 | 0.65 |
| *Correlational* |  |  |  |  |
| Correlation Coefficient () | Cohen (1988) | .10 | .30 | .50 |
|  | Richard et al. (2003)[[4]](#footnote-53)[[5]](#footnote-54) | .10 | .20 | .30 |
|  | Lovakov and Agadullina (2021) | .12 | .24 | .41 |
|  | Paterson et al. (2016) | .12 | .20 | .31 |
|  | Bosco et al. (2015) | .09 | .18 | .26 |
| Cohen’s |  | .02 | .25 | .40 |
| eta-squared () | Cohen (1988) | .01 | .06 | .14 |
| Cohen’s |  |  |  |  |
| Cohen’s f | Cohen (1988) | .10 | .25 | .40 |
| *Categorical* |  |  |  |  |
| Cohen’s omega | Cohen (1988) | 0.10 | 0.30 | 0.50 |
| Phi | Cohen (1988) | .10 | .30 | .50 |
| Cramer’s V |  | [[6]](#footnote-56) |  |  |
| Odds ratio |  |  |  |  |
| Relative risk |  |  |  |  |
| Risk difference |  |  |  |  |
| Cohen’s | Cohen (1988) | 0.2 | 0.5 | 0.8 |

It should be noted that small/medium/large effects do not necessarily mean that they have small/medium/large practical implications (for details see, Coe, 2012; Pogrow, 2019). These benchmarks are more relevant for guiding our expectations. Whether they have practical importance depends on contexts. To assess practical importance, it will always be desirable for standardized effect sizes to be translated to increase/decrease in raw units (or any meaningful units) or a Binomial Effect Size Display (roughly, differences in proportions such as success rate before and after intervention).

**Please also note that only zero means no effect**. An effect of the size .01 is an effect, but a very small (Sawilowsky, 2009), and likely unimportant, one. It makes sense to say that "we failed to find evidence for rejecting the null hypothesis," or "we found evidence for only a small/little/weak-to-no effect" or "we did not find a meaningful effect". **It does not make sense to say, "we found no effect."** Purely by the random nature of our universe, it is hard to imagine that we can obtain a sharp zero-effect result. This is also related to the crud factor, which refers to the idea that "everything correlates with everything else" (Meehl, 1984; Orben & Lakens, 2020, p. 1), but the practical implication of very weak/small correlations between some variables may be limited, and whether the effect is reliably detected depends on statistical power.

## Reporting Effect Sizes

### Transparency

When reporting effect sizes and their calculations, you should prioritize transparency and reproducibility. No matter what tool you used to calculate your effect size (R is the most recommended tool here), you must make sure that others can easily follow your procedures and obtain the same results. This means that if you use online calculators (which is discouraged) or standalone programs (JAMOVI is most recommended; you can also use JASP, which however does not allow access to syntax at this moment), you should include screenshots that capture the input and output, with clear explanations. If you use R, Python or other programming languages, you should copy-and-paste your codes into your supplementary document (or submit your scripts to open online repositories), ideally with annotations and comments explaining the codes. inputs and outputs.

### Directionality

Some effect sizes are directional (e.g., Cohen’s ), which means that they can be positive or negative. Their signs carry important information, and therefore cannot be omitted. When you report these effect sizes, make it clear what is compared to what (i.e., the direction of comparison). Better still, make sure your comparison is inline with the theory. For instance, a theory predicts that your group X should score higher on an item than your Group Y,[[7]](#footnote-59) you should hypothesize accordingly that Group X will have a higher mean than Group Y on the item, and subtract mean(Y) from mean(X) (rather than the other way around) to obtain the mean difference. You should then expect your statistic to be positive, and your value as well. In other words, avoid reporting anything like\*\* = -5.14, = 0.36, where the signs of the statistics do not match.

### Precision

Effect sizes may be very precisely estimated from the available data, the used methodology, and how the population was sampled. It might also be estimated with little confidence on the resulting number. This may be the case for example when the sample is very small, when the population displays a lot of variability, when a between-group design is used instead of a paired-sample design, and finally, when clustered sampling is used instead of randomized sampling. Precision can be estimated using various tools, but probably the most commonly used one is the Confidence intervals. This interval has a confidence level, frequently 95%.

## Interpreting Confidence Intervals

What is the correct interpretation of a confidence interval? Imagine you conducted a study where you compared two groups. You obtained a Cohen’s = 0.3, 95% CI [0.2, 0.4]. How do you interpret this confidence interval?

Confidence intervals are yielded by a certain procedure, such that when the procedure is repeatedly applied to a series of hypothetical datasets drawn from the studied population/populations, it yields intervals that contain the true parameter value (in our example, it means the true difference between the two groups) in 95% of the cases.

In colloquial terms, if we conduct this research over and over (repeating the same sampling procedure, administering the same experimental manipulation, conducting the same statistical analysis, etc.), because of sampling variability (our samples are slightly different at each time), we will get different Cohen’s values. For each of these values, we calculate a 95% interval. Then, among all these many intervals, we expect that 95% of them will contain the true , which we never know exactly.

Therefore, confidence intervals indicate a property of a procedure, not a parameter. One specific confidence interval either contains (the probability is 100%) or does not contain (0%) the true parameter. It is incorrect that we say about any specific CI that “it has 95% probability” of containing that true parameter.

Association, A. P. (2010). *Publication manual of the American psychological association*. American Psychological Association. <https://thuvienso.hoasen.edu.vn/handle/123456789/8327>

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. <https://doi.org/10.1037/a0038047>

Coe, R. (2012). *It’s the effect size, stupid what effect size is and why it is important*. <https://www.semanticscholar.org/paper/It%27s-the-Effect-Size%2C-Stupid-What-effect-size-is-it-Coe/c5ac87df5d6e0e6b6de2f745284835c2a368b0f7>

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Academic Press.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. <https://doi.org/10.1037/a0024338>

Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research: Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. <https://doi.org/10.1177/2515245919847202>

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. *(No Title)*. <https://cir.nii.ac.jp/crid/1130000795088566912>

Harrell, F. (2020). *Author Checklist - data analysis*. <https://discourse.datamethods.org/t/author-checklist/3407>

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. <https://doi.org/10.1037/a0028086>

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. <https://doi.org/10.1002/ejsp.2752>

Meehl, P. E. (1984). Radical behaviorism and mental events: Four methodological queries. *Behavioral and Brain Sciences*, *7*(4), 563–564. <https://doi.org/10.1017/S0140525X00027308>

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. <https://doi.org/10.1177/2515245920917961>

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An Assessment of the Magnitude of Effect Sizes: Evidence From 30 Years of Meta-Analysis in Management. *Journal of Leadership & Organizational Studies*, *23*(1), 66–81. <https://doi.org/10.1177/1548051815614321>

Pogrow, S. (2019). How effect size (practical significance) misleads clinical practice: The case for switching to practical benefit to assess applied research findings. *The American Statistician*, *73*(sup1), 223–234. <https://doi.org/10.1080/00031305.2018.1549101>

Richard, F. D., Bond Jr., C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363. <https://doi.org/10.1037/1089-2680.7.4.331>

Sawilowsky, S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, *8*(2). <https://doi.org/10.22237/jmasm/1257035100>

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00813>

1. The qualification (“almost always”) was that “multiple-degree-of-freedom effect indicators tend to be less useful than effect indicators that decompose multiple degree-of-freedom tests into meaningful one degree-of-freedom effects” (p. 26). “One degree-of-freedom effects” refer to those associated with contrasts, t-tests, F-tests with numerator , and Chi-square tests, whereas “multiple-degree-of-freedom effects” refer to those associated with, for instance, F-tests with numerator , and Chi-square tests with . [↑](#footnote-ref-46)
2. Sawilowsky (2009) expanded Cohen’s benchmarks to include also very small effects ( = 0.01), very large effects ( = 1.20), and huge effects ( = 2.0). It has to be noted that very large and huge effects are very rare in experimental social psychology. [↑](#footnote-ref-51)
3. According to this recent meta-analysis on the effect sizes in social psychology studies, “It is recommended that correlation coefficients of .1, .25, and .40 and Hedges’ (or Cohen’s ) of 0.15, 0.40, and 0.70 should be interpreted as small, medium, and large effects for studies in social psychology. [↑](#footnote-ref-52)
4. “The current review proposes an empirical basis for gauging the size of social psychological effects. It indicates that a correlation coefficient of .10 is ‘small’ relative to most social psychological effects. Mean effects this small are found in roughly 30% of social psychological research literature. It indicates that a correlation coefficient of .20 is a medium-sized effect. Effects that small are found in roughly half of the relevant literature. A correlation coefficient of .30 is large relative to most social psychological effects. Less than 25% of mean effects are that large.” [↑](#footnote-ref-53)
5. These benchmarks are also recommended by Gignac and Szodorai (2016). Funder and Ozer (2019) expanded them to also include very small effects ( = .05) and very large effects ( = .40 or greater). According to them, […] an effect-size r of .05 indicates an effect that is very small for the explanation of single events but potentially consequential in the not-very-long run, an effect-size r of .10 indicates an effect that is still small at the level of single events but potentially more ultimately consequential, an effect-size of .20 indicates a medium effect that is of some explanatory and practical use even in the short run and therefore even more important, and an effect-size of .30 indicates a large effect that is potentially powerful in both the short and the long run. A very large effect size (r = .40 or greater) in the context of psychological research is likely to be a gross overestimate that will rarely be found in a large sample or in a replication.” But see [here](https://twitter.com/aaronjfisher/status/1168252264600883200?s=20) for controversies with this paper. [↑](#footnote-ref-54)
6. The benchmarks for Cramer’s V are dependent on the size of the contingency table on which the effect is calculated. According to Cohen, use benchmarks for phi coefficient divided by the square root of the smaller dimension minus 1. For example, a medium effect for a Cramer’s V from a 4 by 3 table would be .3 / sqrt(3 - 1) = .21. [↑](#footnote-ref-56)
7. Of course, if a theory/effect predicts Group X has a higher mean than Group Y, then it also predicts the reverse, i.e., Group Y has a lower mean than Group X. But theories/effects are commonly articulated in a certain way. It is more common that we say, for example, people prefer the status quo rather than that people do not prefer the non-status quo, when we refer to the status quo bias. Consider another “theory”: teenagers get taller when they get older. It just does not make sense to say the same thing reversely, i.e., teenagers get shorter when they get younger, because people cannot get younger, at least in the 2020s. [↑](#footnote-ref-59)