Effect Sizes and Confidence Intervals Guide

Matthew B. Jané[1], Qinyu Xiao[2], Siu Kit Yeung[3], Daniel J. Dunleavy[4], Lukas Röseler[5], Mahmoud Elsherif[6], Denis Cousineau[7], and Gilad Feldman [8]

[1]Department of Psychological Sciences, University of Connecticut

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

[3]Department of Psychology, University of Hong Kong

[4]College of Social Work, Florida State

[5]University of Bamberg

[6]University of Birmingham

[7]Université d'Ottawa

[8]University 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, 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.

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] to report effect sizes (Association, 2010, p. 34; see Fritz et al., 2012,

[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,

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

---

t-tests, F-tests with numerator $df = 1$, and $1 - df$ Chi-square tests, whereas "multiple-degree-of-freedom effects" refer to those associated with, for instance, F-tests with numerator $df > 1$, and Chi-square tests with $df > 1$.