Metadata of the chapter that will be visualized online

Chapter Title	Effect Size		
Copyright Year	2020		
Copyright Holder	Springer Nature Switzerland AG		
Author	Family Name	**Pardo**	
	Particle		
	Given Name	**Luis**	
	Suffix		
	Division/Department	Department of Psychology	
	Organization/University	University of Chile	AU1
	City	Santiago	
	Country	Chile	AU2
Author	Family Name	**Antivilo**	
	Particle		
	Given Name	**Andrés**	
	Suffix		
	Division/Department	Department of Psychology	
	Organization/University	University of Chile	AU1
	City	Santiago	
	Country	Chile	AU2
Corresponding Author	Family Name	**Miguez**	
	Particle		
	Given Name	**Gonzalo**	
	Suffix		
	Division/Department	Department of Psychology	
	Organization/University	University of Chile	AU1
	City	Santiago	
	Country	Chile	AU2
	Email	gonzalo_miguez@uchile.cl	

E

Effect Size

Luis Pardo, Andrés Antivilo and Gonzalo Miguez
Department of Psychology, University of Chile,
Santiago, Chile

Definition

An Effect Size (ES) is a statistic or a parameter that reflects the quantitative magnitude of a phenomenon or the intensity of an association among two variables in a meaningful way. Furthemore, it is focused on describing the size of the difference or the association rather than confounding this with sample size (Coe 2002). It is precisely for this reason that today there is a clear push for ES to be reported and used for discussing the importance of the results obtained in research studies (Kelley and Preacher 2012). For example, quantifies how fast a rat solves a maze when an experimental training is used vs. a control group.

Introduction

Traditionally, statistical analyses in scientific research have taken the form of significance tests. Fisher developed the Null Hypothesis Significance Testing (NHST), guided by the idea of falsifying the scientist's hypothesis, in line with Popper's falsificationism. Although at first he set the alfa at .05 (i.e., the probability to make a false positive, point out that there is an effect when there is not), later he argued against the idea of using threshold probabilities and pointed out that probabilities can be used as a continuous measure of strength of evidence against the null hypothesis (Fisher 1973); nevertheless, he was mistaken as it does not say anything about the value or the magnitude of that evidence.

Additionally, null hypotheses are almost always false (Kirk 2015), i.e., although trivial, a variable always has an effect over others. Therefore, trivial differences can be labeled as statistically significant with a big sample size and conversely, a limited sample size can output a nonsignificant result, although the difference exists. These lead researchers to control for type I error (reflected in alfa), neglecting type II error (i.e., a false negative, not finding an effect that exists; Kirk 2015).

Finally, Null Hypothesis Significance Testing approach goes in another direction than scientific inference (Kirk 2015). The first one is concerned with the probability of obtaining a set of data (or more extreme) if the null hypothesis is true, while the latter has to do with the probability that the null hypothesis is correct based on available data. These two probabilities rarely coincide (Falk 1998).

Criticisms about the Null Hypothesis Significance Testing have led quantitative psychologists to search manners to complement it. One of them is the effect size. Despite the successive calls from the American Psychological Association (APA;

J. Vonk, T. K. Shackelford (eds.), *Encyclopedia of Animal Cognition and Behavior*,
https://doi.org/10.1007/978-3-319-47829-6_242-1

e.g., APA 2020) encouraging its use, the reporting of ES keeps being limited (Barry et al. 2016).

In the literature, there are several definitions of effect size (we presented a general one above). For instance, for Kelley and Preacher (2012) is a quantitative reflection of the magnitude of a phenomenon used for addressing a question while for the APA and the AERA (American Educational Research Association) is a population-based parameter that can be estimated from samples AU4 with uncertainty and reflects the index of the effect or the relation between variables (Peng and Chen 2013). According to Kirk (2015), it can be a statistic or a parameter that allows to quantify the size of a phenomenon of interest in an interpretable way. As can be noticed, APA and AERA definition emphasizes the function while others the statistic to be used. Nevertheless, all of them point to the same idea.

Commonly, it has been pointed out that ES are free-scale measures. That is not totally true because some are standardized while others don't. Using a standardized ES allows you to compare data from different scales of the same conceptual variable. The use of an effect in the original metric is easier to interpret.

As has been noted above, hypothesis testing is very sensitive to the sample size. In contrast, ES does not depend on the sample size; however, it can be better estimated with bigger sample sizes.

ESs are very useful in behavioral science. Besides its uses as a report of the quantitative magnitude of a phenomenon, it also allows calculating the required sample size to achieve an acceptable power (i.e., the capacity of the study's to find the phenomenon, inversely related to type II error), integrate the literature from different experiments and reports about the same question by meta-analytic techniques, and determine the practical significance of research results.

Effect Size Estimates

There are a lot of effect size measures and most of them can be grouped in two categories: differences between groups (the d family) and measures of association (the r family).

Difference Between Groups

The procedures that belong to this family estimate difference of parameters like differences in means or proportions. The group's comparison can be done on continuous and dichotomous variables.

Continuous Data

The first effect size measure explicitly labeled as such is part of this family and was introduced by Cohen (1969) is the parameter δ, given by:

$$\delta = \frac{\mu_E - \mu_C}{\sigma}$$

where μ_E stands for the population mean of experimental group and μ_c is the population mean of the control group. σ represents the common population standard deviation. Given that the numerator can be influenced by the scale of measure of the means, is divided by σ to rescale it in units of the amount of error variability in the data (Kirk 2015). For that reason, it is recognized as a standardized measure.

To estimate δ, Cohen's d, Glass's g, and Hedges's g had been developed. They have in common that are standardized ES estimators and assume normality of the distribution of data and homoscedasticity (i.e., similar variance between the groups; Peng and Chen 2013; Kirk 2015).

Of these, Cohen's d is perhaps the most commonly reported. It ranges from negative infinity to infinite, where the sign – positive or negative – indicates the direction of difference between groups. It is calculated as follows: AU5

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sigma_{\text{pooled}}}$$

where $\overline{\mathrm{X}}$ stands for sample mean of each group and σ_{pooled} is calculated as:

$$\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2}}$$

With n= sample size and s^2 is the variance statistic.

Cohen (1988) proposed that Cohen's d of 0.2 is a small effect size, while 0.5 and 0.8 are medium and large, respectively. However, an effect size must be interpreted as small or large linked to the research area, the methods employed and the outcome.

Despite its widespread use, Cohen's d has some problems. In the first place, it is a biased estimator of δ. Also, the estimation of σ depends on the groups size. In the third place, it must be said that if the population variances are not equal, it is problematic estimate δ from Cohen's d because the parameter itself is undefined. Finally, this estimator is sensitive to outliers.

Another estimator of δ is Hedges g. It only introduces a little change in the way of estimating σ_{pooled} respect to the formula just presented: to the sum of group sample sizes is subtracted the value of two (i.e., n1+ n2−2; Hedges 1981)

Finally, we present Glass's g, which was developed for meta-analysis in the context of an experimental study with one control group and an experimental one. Glass reasoned that comparisons of different experimental groups with control could lead to differing estimates of σ_{pooled}. Due to this, instead of dividing by σ_{pooled} as in Cohen's d, the mean difference is divided by control's group sample standard deviation (Glass 1976). Depending on the research design, it could be estimating different parameters, so researchers must be cautious. It is traditionally recommended when the homoscedasticity assumption is violated.

Dichotomous Variables

When the groups are compared on two-level variables (e.g., solve vs. fail to solve a recognition task), comparisons can be based on the probability that the members of a group will be classified in one of the categories.

The two measures that we present here, risk ratio and odds ratio, compare how possible is the occurrence of an event in one group relative to the other group. The first one understands likelihood in terms of probabilities while the second does it in terms of odds. These indexes will be explained using a fictional example.

Imagine that a group of researchers conduct an experiment to learn if the maternal consumption of THC affects the recognition ability of its descendants (measured in an object recognition task, widely used in the study of rodentia memory and the study of memory and learning other species). They obtain the following results, expressed as row data (probabilities) (Table 1):

The risk ratio is calculated dividing the probability of the outcome in one group (p) over the probability of that outcome in the other group. In this example, the outcome is the number of rats that solved the task and the groups are THC (i.e., p) and No-THC (i.e., q). In this case it is 4.5.

To calculate the odds ratio, the other measure of this group presented here is needed to compare the odds of certain outcome for each group. The formula is as follows:

$$\mathrm{OR} = (p/(1-p))/(q/(1-q))$$

Replacing the values with the above-mentioned outcome, we have:

$$\mathrm{OR} = (0.9/(0.1))/(0.2/(0.8)) \Rightarrow 36$$

If the result of this division were smaller than one, then the outcome is more likely in the second group. If it is one, then there is no difference. If it were larger than one, then the outcome is less likely in the second group. In this case, the odds ratio is 36, which means that the likelihood of give birth to live born rats is minor in the group of rats treated with THC.

Association Measures

The ES of this family includes measures of association among two or more variables.

Pearson's r (developed in 1896) is probably the most known and several of the other members of this groups are variations of r. Indicates the degree of linear association between two variables and ranges from -1 to $+1$, where $|1|$ means perfect association. It must be taken account that the scale is not linear (e.g., 0.2 is not twice 0.1). When it is squared, we have the coefficient of determination used in bivariate regression analysis and indicates how much of the variance of one variable can be

Effect Size, Table 1 Caption

	Solve the task	Fail to solve the task	Total
THC	27 (0.9)	3 (0.1)	30
No-THC	7 (0.2)	28 (0.8)	35

Note:

attributed to the other. It is measured in terms of percentages.

In the context of multiple regression analysis, when one variable can depend on a set of predictors, it is used the coefficient of multiple determination (R^2). However, it can be inflated through the sample size and the number of predictors in the model. One alternative developed to deal with this is the adjusted R^2.

Eta^2 or correlation ratio also can correct the resulting inflation of R^2, although partially. It is more common in psychology that R^2, allowing to know how many of the dependent's variable variance is accounted by the group membership of the subjects. *Eta*^{2} is also associated to – one-way – Analysis of Variance (ANOVA; i.e., there are more than two groups).

Finally, we present the Cohen's f. It is used in the context of multiple regression but also in ANOVAS as *Eta*^{2}. In the last case, it can be considered as an extended version of Cohen's d and it is used to measure the dispersion of means among three groups or more.

Reporting and Interpreting ES

Although standardized effect size measure makes easier the processes of meta-analysis and power analysis (for example, Cohen's d can be transformed to Pearson's r), the literature suggest the use of ES in the original metrics of the variables in the context of primary research. Both may be useful.

It may be obvious, but the ES reported must be specified. Also, given the lack of agreement on names for estimates of δ, for example, its suggested to inform how ES was calculated (Appelbaum et al. 2018).

Additionally, it is recommended to provide a confidence interval (Steiger 2004). This allows to quantify the accuracy of the point estimate. This must be considered when interpreting. The wider the confidence interval, less mature is a research paradigm. Thus, the research and historical context must be taken account in the interpretation.

Cross-References

- ▶ Fisher
- ▶ Odds Ratio
- ▶ Rodentia Memory

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3. https://doi.org/10.1037/amp0000389.

Barry, A. E., Szucs, L. E., Reyes, J. V., Ji, Q., Wilson, K. L., & Thompson, B. (2016). Failure to report effect sizes. *Health Education & Behavior, 43*(5), 518–527. https://doi.org/10.1177/1090198116669521.

Coe, R. (2002, September 12–14). *It's the effect size, stupid: What effect size is and why it is important*. In [Conference presentation] Annual Conference of the British Educational Research Association. 2002 Exeter, England. http://www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Falk, R. (1998). Replication—a step in the right direction. *Theory & Psychology, 8*, 313–321.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.3102/0013189x005010003.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128. https://doi.org/10.3102/10769986006002107.

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

Kirk, R. E. (2015). Effect size measures. *Wiley StatsRef: Statistics Reference Online*, 1–13. https://doi.org/10.1002/9781118445112.stat06242.pub2.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philosophical Transactions of the Royal Society of London, 187*, 253–318. https://doi.org/10.1098/rsta.1896.0007.

Peng, C.-Y. J., & Chen, L.-T. (2013). Beyond Cohen's d: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education, 82*(1), 22–50. https://doi.org/10.1080/00220973.2012.745471.

Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*(2), 164–182. https://doi.org/10.1037/1082-989X.9.2.164.

Author Queries

Chapter No.: 242-1 418993_0_En

Query Refs.	Details Required	Author's response
AU1	Please confirm the affiliation detail is fine.	
AU2	Please be aware that your name and affiliation and if applicable those of you co-author(s) will be published as presented in this proof. If you want to make any changes, please correct the details now. Note that corrections after publication will no longer be possible. If no changes are required, please respond with "Ok".	
AU3	Reference Fisher (1973) is cited in the text, but not in the list. Please provide bibliographic details for it or delete this citation from the text.	
AU4	The citations Peng and Chen (2014), Pearson (1886) have been changed as Peng and Chen (2013), Pearson (1896).	
AU5	Please confirm the math representation for "**Direction of Difference**" is okay.	
AU6	Please check the phrase "to the sum of group sample sizes is subtracted the value of two" for clarity.	
AU7	Please confirm the call-out for Table 1 is fine.	
AU8	Please provide caption and table-note for Table 1.	
AU9	Please check if edit to the text "Replacing the values with the above-mentioned outcome, we have:" is okay.	

Note:
If you are using material from other works please make sure that you have obtained the necessary permission from the copyright holders and that references to the original publications are included.