Effect size

Marie Delacre[1], Christophe Leys[1], Limin Liu[2], & Daniël Lakens[3]

[1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

[2] Université de Gant

[3] Eindhoven University of Technology, Human Technology Interaction Group, Eindhoven, the Netherlands

Author Note

Correspondence concerning this article should be addressed to Marie Delacre, CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Abstract

Keywords: keywords

Word count: X

Effect size

Intro

During decades, researchers in social science (Henson, 2000) and education (Fan, 2001) have overestimated the ability of the null hypothesis (H0) testing to determine the importance of their results. The standard for researchers in social science is to define H0 as the absence of effect (Meehl, 1990). For example, when comparing the mean of two groups, researchers commonly test the H0 that there is no mean differences between groups (Steyn, 2000). Any effect that is significantly different from zero will be seen as sole support for a theory. Such an approach has faced many criticisms among which the most relevant to our concern is that the null hypothesis testing highly depends on sample size: for a given alpha level and a given difference between groups, the larger the sample size, the higher to probability of rejecting H0 (Fan, 2001, Sullivan_Feinn_2012, Olejnik_Algina_2000, Kirk_2009). It implies that even tiny differences could be detected as statistically significant with very large sample sizes [McBride, Loftis & Adkins, 1993]. This is especially problematic since, a zero effect rely on the assumption that subjects are randomly assigned to conditions, and even if so, there are many circumstances in which the probability of getting an actual zero effect is very low. For example, Meehl (1990) argues that factors of no theoretical interest might influence results because they are interaction between some of these factors and the manipulated variable (for example, when testing a placebo against a anti-depressor, a given ingested food could influence the effect of the anti-depressor in a different way than it would influence the placebo effect).

Facing these arguments, it has now become common practice to report the *p*-value assorted by a measure of the effect size, that is, a quantitative measure of the magnitude of the experimenter effect. Traditionnaly, most researchers report a few common measures, namely the cohen's *d* when using *t*-tests, the eta-square (or partial eta-square) when using ANOVA or factorial ANOVA and the R-square when using correlations or regressions #JE

NE SAIS PAS SI TU ES D'ACCORD AVEC CA, FAUDRAIT PEUT ETRE UNE REFERENCE QUI EN PARLE#. However, we will argue that in the current litterature, several concepts address the problem of p-value and effect size at different level of analysis, which introduces a lot of confusion. Moreover, there are several situations that call for effect size measures. Following the nature of these situations, each indicator of effect size has its strenghts and weaknesses. Lastly, each effect size is submitted to a range of assumptions that are unrealistic in many research designs. As consequences many effect size estimators are inaccurate and alter the robustness of the statistical conclusions.

In sum the aim of this paper is fourfold: (1) Discuss the different levels of significance, following Bothe and Richardson (2011)'s nomenclature, and distinguish their domain of validity and explain the stakes related to effect size.(2) Identify the different situations that justify the use of effect size. (3) Identify the strenghts and weaknesses of several measures of effect sizes in these situations; (4) Discuss the problem of the effect size assumptions and the impact of these assumptions violations on the robstness of the measures of effect size and the statistical conclusions.#JE NE SUIS PAS ENCORE TOUT A FAIT CLAIR AVEC CETTE STRUCTURE, ON POURRA LA FAIRE EVOLUER APRES#

#Statistical, Practical and Clinical Significance

In their paper, Bothe and Richardson (2011) distinguish between three levels of significance, namely Statistical significance, Practical significance and Clinical significance (with the adjunction of Personal significance). Statistical significance refers to the p-value. As stated before, this conclusion is highly dependent from the sample size. Practical significance refers to the magnitude of a change or a difference between groups. In other terms, it is any statistical indicator that assess mathematically the effect size. Laslty, the Clinical significance refers to the interpretation of treatment outcomes. This last level is not statistical nor mathematical, it is related to underlying theory that posits an empirical hypothesis.

It is important to understand the difference between these three concepts. Statistical significance allows the researcher to determine whether the oberved departure from H0 can be attributed to something else than randomness (i.e. an actual effect). Practical significance is a mathematical indicator of effect size that is not necessarily related to the theoretical effect or at least that the relation is not straightforward. As stated by Kazdin « … clinical significance has been defined as whether an intervention "makes a real (e.g., genuine, noticeable) difference in everyday life to the clients or to others with whom the clients interact"" (Kazdin, 1999, p.332 ; cité par Bothe & Rochardson, 2011). Indeed, Pratical significance depends on the way a variable is converted into numerical indicator. For example, when assessing Self-Compassion, one can use a scale such as the Self-Compassion Scale (Kotsou & Leys, 2016; Neff, 2003) #Kotsou, I., & Leys, C. (2016). Self-Compassion Scale (SCS): Psychometric properties of the French translation and its relations with psychological well-being, affect and depression. PloS one, 11(4), e0152880.; Neff, K. D. (2003). The development and validation of a scale to measure self-compassion. Self and identity, 2(3), 223-250.#. This scale inform the researcher about the level of self-compassion based on a ordinal scale that can yield different value depending on the influence of any independent variable. For example, some training program can improve subjects level of self-compassion (Jazaieri et al., 2013) #Jazaieri, H., Jinpa, G. T., McGonigal, K., Rosenberg, E. L., Finkelstein, J., Simon-Thomas, E., … & Goldin, P. R. (2013). Enhancing compassion: A randomized controlled trial of a compassion cultivation training program. Journal of Happiness Studies, 14(4), 1113-1126.#. Yet, since the scale is ordinal, meaning that there is no standard unit to assess the construct, the relation between the mathematical effect size (i.e. Pratical significance) and the actual change in self-compassion (i.e. Clinical significance) will always remain unknown. Therefore, although, as we will see, pratcial significance is important to determine, it's relation with clinical significance has often to be addressed, and that is more a theoretical argument than a statistical one. Lastly, even on stronger scales of measurement effect size can be of little clinical value. For example, Ogles and Bonesteel

(2001) describe the case of a diet that will allow overweighted people to loose 16 pounds. In some cases people will still be overweighted and suffer from physical incomfort or obesity related diseases where as for others, the same effect size could lower the weight to the point were physical limitations disappear and obesity related diseases are cured. In the first case, results are of lesser clinical significance than in the second case, although mathematically they are of comparable practical significance. These three levels of significance are of course positively correlated. Indeed, a large effect size will probably not appear randomly (hence will have a small *p*-value) and will more often than not be of theoretical interest. However they can act independently as well. For examples, in some situations, if the sample size is small, even large effect sizes could still be consistent with H0 where as in other situations even small effect sizes can be of theoretical value (and be statistically significant provinding a sufficient sample size). It is therefore important to disentangle these three levels of significance while arguing about statistical conclusions.

#What are the four situations that call for an effect size estimator?

and we will focus on two of them. First, there are discussions related with the definition of the null hypothesis itself. A (REMARQUE: Due to this factors, one could observe an effect which is reliable but not of interest, et JE NE SUIS PAS SURE QUE LA TAILLE D EFFET REPONDE A CE GENRE DE SOUCI). This is related with what is probably the most famous criticism of . AS a conclusion, *statistically significant* effect is not necessarily of *practical* interest. The *statistical* significance is the probability that findings have occured by chance (Stout, 1995). The *practical* significance is the magnitude of findings and is assessed by measures of **effect sizes**. Reporting measures of effect size (and a confidence interval around this measure) has been recommended for more than 50 years [Fan (2001); Hays (1963);Cohen 1965] and is highly encouraged by the *APA Publication Manual* [APA, 2010], even if it does not seem to have fully entered the mores. Attention: les deux sont complémentaires et donc ES ne doit pas “remplacer” mais “compléter”.

At the same time, a vast literature has developed that casts doubt on the credibility of the assumptions of Student's t-test and classical F-test ANOVA (i.e. the assumptions that two or more samples are independent, and that independent and identically distributed residuals are normal and have equal variances between groups; Glass, Peckham, & Sanders, 1972) (CITER TOUTES MES REFERENCES). In a previous paper, We focused on the assumptions of normality and equality of variances, and argued that these assumptions are often unrealistic in the field of psychology. Bcp d'autres chercheurs avant nous étaient arrivés à la même conclusion. Pourtant, beaucoup moins d'auteurs se sont penchés sur les mesures de taille d'effet à utiliser en complément du test de welch. Il existe de la littérature sur la question, mais pas vraiment d'accord (parce que grande confusion quant à la questino suivante: à quoi sert la mesure de taille d'effet?) Par ailleurs, s'il est de plus en plus communément admis que les conditions d'application des tests de comparaison de moyennes (dominant toujours la recherche) sont peu réalistes et rarement respectées, pourtant et que de nombreux chercheurs recommandent d'utiliser le Welch au lieu du test de Student, peu de littérature suggère quelle taille d'effet associer à ce test. Même Jamovi ne propose comme mesure de taille d'effet que le d de Cohen, souffrant des mêmes limites que le test de Student.

Pour cette raison, nous proposons de structurer cet article comme suit: # 1) Bien definir practical significance (donc donner une définition claire de la taille d'effet qui nous convient) Expliquer un peu pourquoi c'est important d'avoir l'IC autour de l'effect size: 1) Parce que l'estimation dépend du n (plus n est grand, plus précise est l'estimation) 2) parce que la mesure de taille d'effet est un complément de la significativité statistique: comme le dit

2) Bien définir à quel objectif on tente de répondre via la mesure de taille d'effet (je les cite tous dans mon pwp)

3) Qualités MATHEMATISUES importantes d'une bonne mesure de taille d'effet et de l'IC

4) Revue sur les familles de tailles d'effet (r et d, et mesures les plus connues)

5) Simulations

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*(5), 275–282. doi:10.1080/00220670109598763

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

Hays, W. L. (1963). *Statistics for psychologists* (Holt, Rinehart & Winston.). New York.

Henson, S., R. I. (2000). State of the art in statistical significance and effect size reporting: A review of the APA task force report and current trends. *Journal of Research and Development in Education*, *33*(4), 285–296.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Steyn, H. S. (2000). Practical significance of the difference in means. *Journal of Industrial Psychology*, *26*(3), 1–3.

Stout, R., D. D. (1995). Assessing the practical signficance of empirical results in accounting education research: The use of effect size information. *Journal of Accounting Education*, *13*(3), 281–298.