

Effect size

Marie Delacre¹, Christophe Leys¹, Limin Liu², & Daniël Lakens³

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

10

Abstract

11

12 *Keywords:* keywords

13 Word count: X

Effect size

Intro

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.

During decades, researchers in social science (Henson, 2000) and education (Fan, 2001) have overestimated the ability of the null hypothesis testing to determine the importance of their results. The standard for researchers in social science is to define the absence of effect as the null hypothesis (Meehl, 1990). For example, when comparing the mean of two groups, researchers commonly test the null hypothesis that there is no mean differences between groups (Steyn, 2000). Any effect that is significantly different from zero will be seen as support for a theory. Such an approach has faced many criticisms and we will focus on two of them. First, there are discussions related with the definition of the null hypothesis itself.

A zero effect is possibly meaningful only if subjects are randomly assigned to conditions, and even if so, there are many circumstances in which the probability of getting a zero effect is very low. For example, Meehl (1990) argues that there are often systematic factors in action, i.e. factors that we are not theoretically interested in and that influences results. (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). Second, even when there are no differences between groups in the population, it is very unlikely to observe no differences between sample groups, due to sampling error (Prentice, 1990). This is related with what is probably the most famous criticism of the null hypothesis testing: it highly depends on sample size: for a fixed alpha level and a fixed difference between groups, the bigger the sample size, the higher to probability of rejecting the null hypothesis (Fan, 2001, Sullivan_Feinn_2012, Olejnik_Algina_2000, Kirk_2009). It means that even tiny differences could be detected as statistically significant with very large sample sizes [McBride, Loftis & Adkins, 1993]. 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”.

Pour cette raison, nous proposons de structurer cet article comme suit: # 1) Bien définir 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

(REMARQUE: QLQ PART DANS L'ARTICLE, IL FAUDRA QD MM BIEN RAPPELER QUE SIGNIFICATIVITE PRATIQUE n'est pas toujours synonyme de SIGNIFICATIVITE clinique... si notre but est de clarifier les effect size, il faut être bien clair nous-mêmes). The *practical* significance is the magnitude of the effect.

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.

Prentice, M., D.A. (1990). When small effects are impressive. *Psychological Bulletin*,

90 112(1), 160–164.

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

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