Vol.8.Issue.4.2020 (Oct-Dec)

http://www.bomsr.com
Email:editorbomsr@gmail.com

REVIEW ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

A COMPREHENSIVE INTRODUCTION TO EFFECT SIZE IN ASSOCIATION WITH NULL HYPOTHESIS SIGNIFICANCE TESTING

CHUDA PRASAD DHAKAL, *PhD*
Tribhuvan University, Institute of Agriculture and Animal Sciences / Post Graduate Campus Kirtipur, Kathmandu, Nepal
chudadhakal@iaas.tu.edu.np
DOI: 10.33329/bomsr.8.4.50

ABSTRACT

The techniques of null hypothesis significance testing are diminishing in use whereas there has been an increasing trend of seeking practical significance i.e. applying of effect size approach in them. This article aims to comprehend effect size technique in association to null hypothesis significance testing. Prevalent literature on effect size along with null hypothesis significance testing was extensively reviewed and synthesized. Null hypothesis significance testing in isolation is reported not sufficient but if combined arduously with effect size much more practically conclusive and meaningful research findings are possible. Application of effect size and its confidence interval are essential to give any research a better meaning in real life. And the compulsory reporting of effect size in scholarly journals has led to envision that in the near future, researchers would seek their research findings to be not only likely but also as practically significant as possible.

Keywords: null hypothesis significance testing; effect size; cohen's d; confidence interval; practical significance

1. BACKGROUND

According to Gigerenzer & Murray (as cited in Levine et al., 2008), the first basic significance test that came across developing since 1900 was founded back in 1710. Pearson developed the first modern significance test (the chi-square goodness-of-fit test) in 1900. Soon after this Gosset published his work leading to the development of the t-test in 1908. With time, Null Hypothesis

Significance Testing, here after abbreviated as NHST, became a widely accepted and frequently used statistical framework to provide evidence of an effect. To date, NHST has been a widely applied aid to the interpretation of experimental data in the study of numerous disciplines.

Despite its wide and popular use, NHST has plenty of criticisms and debates about its correct application and validity in the scientific studies. A couple of studies that adhere the shortcoming, insufficiency or misconception about NHST are [(Levine et al, 2008); (Denes & Ioannidis, 2015); (WikiVet, 2011); (Null hypothesis significance testing, n.d.); (Limitations of significance testing, 2015); (Szucs and Ioannidis, 2017); (Castillo & Torquato, 2018)]. These studies have mentioned the controversies and the problems with the approach of NHST which primarily suggest taking utmost care while interpreting the results in the due course of its application. Always interpret the parameter estimates and effect sizes as well as *p*-values (Field, 2020).

In another perspective, NHST is the only decision-making mechanism so far. For this reason, scientists are compelled to use it despite its limitations. To make intelligent decisions NHST techniques should be combined with other techniques keeping a clear insight into their underlying concepts along with their limitations and the proper interpretation of statistical evidence. Such improvements are then justified in the case of pre-study power calculations. A couple of studies which argue this notion are: [(Nickerson, 2000); (Hypothesis Testing: Methodology and Limitations, 2001); (Wikivet, 2011); (Limitations of significance testing; 2015); (Null hypothesis significance testing, n.d.); (Pernet, 2016)].

In the context of susceptibility of NHST, Szucs and Ioannidis (2017) and Huberty (2002) suggest NHST should no longer be the default method for decision making but the effect size that fills in the blank as its best alternative. These studies argue that limitations of NHST are overcome by using effect size in combination with hypothesis test *p*-value.

As NHST techniques are diminishing in use. Whereas there has been a rise of effect size approach. This article thus is, a comprehensive review of effect size techniques. This includes demonstrating the vitality, computation and interpretation of effect size for quality research. However, it is always limited to clarifying the concepts and the constructs it aims to disseminate.

2. WHAT IS EFFECT SIZE

Effect size is primarily to displace the NHST techniques because NHST has plenty of ambiguities. For instance, NHST is not sufficient to make intelligent decision. It is not a complete test of significance as the only take-home message it gives is either or not the treatment has an effect. But it does not tell how big or small the effect is to a sample. I.e. what amount of variation does the treatment bring to a sample is not answered. As a result, should the effect be considered in real practice or not is not clear. Contextually claimed by Sullivan and Feinn (2012), is : 'statistical significance is either questioned or is ought to be insightfully employed.'

Huberty (2002) has mentioned that history of effect size started discussing (a) relationship, (b) group differences, and (c) group overlap, at around 1940. It is found that since last several decades, in most of the studies, emphasis is given on the reporting and interpretation of effect sizes. Statistics how to (2020) literaly defines effect size as the measurable variation that a treatment brings to a sample.

A p-value included in statistical research tells which treatment, methodology or any other intervention is statistically sounder than its alternative. But it is not about the quantity by how much the effect is. This (p-value), therefore does not have any practical significance. Danial (2017) argues that in NHST there is nothing magical about p =.05. This (p =.05) does not have any meaning in conclusion and only signifies if a treatment has an effect. But it cannot say anything on whether to consider that effect or not. Thus, it is not sufficient for further action, that the groups are different.

Hence, effect size is not the same as statistical significance. It is about size of the effect of a treatment to a sample. Effect size tells how important the result is, but statistical significance is how likely is that a result is due to chance. In a *statement on statistical significance and P-values*, Wasserstein & Lazar, (2016, p.132) explains:

> Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even a lack of effect. An effect size is a measure of how important a difference is: large effect sizes mean the difference is important; small effect sizes mean the difference is unimportant. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

Statistical significance is limited to only signifying the groups are different based on obtained *p*-value but nothing afterwards. In this context [Sullivan & Feinn (2012); (Effect size, 2020); (Effect Size, 2019); (Effect size: What is it and when and how should I use it, n.d.)] have mentioned that effect size fills in the blank. That is, as a quantitative measure of the magnitude of a phenomenon or as the size of the effect of a treatment or as a distance that a treatment pushes a sample away from means or as a concept that measures the strength of the relationship between two variables on a numerical scale, effect size enables to see how substantially different the prior and the posterior results are due to treatments.

For example, when a treatment is employed to a sample effect size shows how much it pushes the sample from the position where it was before employing the treatment. In such way statistical effect size helps in determining if the difference is real. And if the size of the difference is real then it matters. Effect size is impact of a finding and it is what we refer to as practical significance. Also, effect size is to seek why any size of the effect of treatment was not statistically significant in the case where a non-significant result was observed.

For this reason, arguments are found that explains that in any studies effect size and the specific *p* values should be reported to make the findings more meaningful. Statistical significance is not comparable. It is either significant or non-significant; nothing else. But in practice, even the non-significant results can be important as statistical significance is not the same as practical significance.

Moreover, Daniel (2017) in his YouTube Video files entitled, 1) “What Statistical Significance Means – Part 1 (8-11)”, 2) “Cohen’s d effect size for t-tests (10-7)” and 3) “What Statistical Significance Really Means (10-6)”; have summarized the issue as:

> Statistical significance tells us that the differences that we found are unlikely to be due to chance or luck, but possibly not. And even if the differences are real, they may not mean

anything in the real world. And that is why, whenever we run a statistical test, we should also compute a measure of effect size.

Accordingly, in his you tube video entitled "Cohen's d effect size for t-tests (10-7)" the author has defined effect size as "a standardized measure of the size of an effect which can be objectively compared to determine whether the treatment had any practical usefulness."

The end point therefore is, effect size is the measurable strength or impact of a finding which conveys the message how important a treatment is instead of only leaving a comment, whether the treatment makes a difference to a sample or not as NHST which won't have any meaning in conclusion. But effect size can determine whether the treatment had any practical usefulness to consider it in real.

3. WHY EFFECT SIZE

The difference between the groups investigated by NHST techniques is meaningless in the real world. They do not help readers understand the magnitude of differences found in the studies. In an experiment, if a treatment has an effect greater than zero researchers want to know how big the effect is. Such measures of the difference between the groups help readers understand the importance of their findings. It will make them able to decide either or not to consider the treatment in real. Field (2020) reveals that we can go beyond p value to evaluate the plausibility of a hypothesis, effect sizes that address more useful questions, are less dependent on sample sizes; they quantify the size of the effect and encourage thinking about effects on a continuum. Daniel (2017) insists to include some index of effect size or strength of the relationship in the results section and Sullivan & Feinn (2012) recommends reporting both the substantive significance (effect size) and statistical significance (*p*-value) in the study report for readers to understand the full impact of their studies.

Reporting effect sizes are considered good practice when presenting empirical research findings in several disciplines. It facilitates the interpretation of the functional as opposed to the statistical significance of a research result. And this is the reason why effect size in the research reports has almost been a must factor to report. American Psychological Association (APA) Task Force on Statistical Inference recently emphasized, "Always provide some effect size estimate when reporting a *p-value*" (Wilkinson & APA Task Force on Statistical Inference, 1999, p. 599). (as cited in Huberty, 2002).

According to Lakens (2013) & Neill (2008) effect sizes complement statistical hypothesis testing and play an important role in power analyses and sample size planning. The magnitude of the reported effects in standardized metrics help communicate the practical significance of the results that the researchers have found and provide context to the event. For instance, when the sample size is small, effect sizes give meaning and when the sample size is large, effect sizes lend clarity. Meta-analytic conclusions are made possible by comparing standardized effect sizes across studies. Effect sizes from previous studies can be used when planning a new study. The S.E. of the effect size is used to weigh effect sizes when combining studies, so that large studies are considered more important than small studies in the analysis.

Sullivan & Feinn (2012) argues, an estimate of the effect size is often needed before starting the research endeavor in order to calculate the number of subjects likely to be required to avoid a Type II error (the probability of concluding there is no effect when one actually exists). This explains as, one must determine what number of subjects in the study will be sufficient to ensure (to a

degree of certainty) that the study has acceptable power to support the null hypothesis. That is, if no difference is found between the groups, then this is a true finding. Further, for the purpose of calculating reasonable sample size, the effect size can be estimated by pilot study results, similar work published by others, or the minimum difference that would be considered important by educators/experts.

4. HOW TO EFFECT SIZE

Practically every study which is capable of being verified or disproved by observation or experiment looks for an effect. Effect size is the amount of anything that is of research interest (Stukas & Cumming, 2014). This quantifies the magnitude of the effect that emerges from the sampled data (i.e., the difference between populations or the relationship between explanatory and response variables).

Importance of effect size goes beyond its mandatory reporting in the research reports. Effect size basically measures practical difference between the groups or the relationship between variables. However, regarding the deeper understanding and application of effect size Durlak (2009) explains, 'an effect of lower magnitude on one outcome can be more important than an effect of higher magnitude on another outcome.' Another application of effect-size is as a standardized index. It is independent of sample size. Also effect sizes are used in power analysis.

Interpreting effect size

Effect sizes may be in original units, or units free or they may be standardized or in a squared measure. Different types of effect size serve for different purposes. Martin (2020) explains, there are two types of effect size. Simple effect size and standardized effect size. Simple effect sizes are one which describe the size of the effect with the original units of the variables. But standardized effect sizes are unit free (size of the effects in the concerned studies are divided by the relevant standard deviations) and their interpretation ultimately leads to the question of what is a small, medium, or large effect. Most often standardized effect sizes are used. One typical benefit of standardized effect size is that it makes possible to compare the results across studies. However, in situations where the difference would be better expressed with the original units of the variables, simple effect size is used. Recommendation made is to present effect sizes always for primary outcomes.

According to (Stukas & Cumming, 2014) interpreting effect sizes is a challenging task. To produce more conclusive evidence in the research, calculating, reporting, and discussing effect sizes should be highly valued. It needs to acknowledge the uncertainty in an effect size estimate as in confidence interval. The advantages of confidence intervals are that they facilitate data interpretation and easily detect trivial effects (Téllez et al., 2015). A confidence interval with a 95 % confidence level has a 95 % chance of capturing the population mean. Technically, this means that, if the experiment were repeated many times, 95 % of the CI would contain the true population mean. As confidence interval allows readers to assign practical meaning to the values it is therefore strongly recommended to report.

Besides, a broader coverage while interpreting effect size consists of incorporating independent variables, participants, comparison with other results in the research field etc. Further, Kirk (1996) (as cited in LeCory and Krysik, 2007) argues that there are little point presenting effect sizes in papers if these are not interpreted and discussed correctly. For instance, effect size calculated from two variables is appropriate if there are no influential covariates and the sample size issue is correctly dealt. Also, problems with heterogeneous data and non-independence of data need

to be considered. Lastly, translating effect size into practical importance is always essential and practicable and wherever that can meet research goals and help meaningful interpretation, simple rather than complex effect size is recommended.

Cohen's d effect size

While talking about effect size, firstly, appropriate effect size measure is to be traced out for any study. Here we have considered *Cohen's d effect size* to discuss and present in detail. If two groups of the same size have similar standard deviations Effect Size Calculator for T-Test (2020) has mentioned *Cohen's d* is the appropriate effect size measure. According to which, for the independent samples t-test, *Cohen's d* is determined by calculating the mean difference between the two groups, and then dividing the result by the pooled standard deviation.

Cohen's d therefore is a standardized effect size. This is unit free measure and can compare the results across studies. For instance, if two conditions mean length are 2.3cm and 1cm, the simple effect size would be the difference in the mean length i.e.1.3 cm. This is the best estimate of the difference, the point estimate. For which with 95% confidence the difference in the means comes to be [between 0.97cm and 1.63 cm]. This is the range of the values for the difference [we estimated this arbitrarily designating the values for pooled standard deviation is 1, and sample size for the first group is 75, and that for the second group is 70]. Where the center of the confidence interval (the mean difference) is the most reasonable value and the ends are less plausible values for the population mean difference. But for the same, standardized effect size would have been 1.3. Accordingly, to be interpreted in terms of standard deviations.

In addition, if there would be more than two group means, *Cohen's d* effect size measure would be the difference between the largest and smallest means divided by the square root of the mean square error. However according to Martin (2020) if each group has a different standard deviation appropriate effect size measure would be *Glass's delta* which uses only the standard deviation of the control group. But if the sizes of the two samples are different, then *Hedges' g* effect size is used.

Few other effect sizes mentioned in 'How is the effect size used in power analysis (2020) are: *F-ratio* effect size used for the regression coefficient in a regression analysis and in analysis of variance, *Pearson r* effect size for correlation between two variables and χ^2 - effect size (the best statistic to measure the effect size for nominal data) for contingency tables.

According to Cohen (1988) [as cited in 'how is effect size used in power analysis' (2020)] for low, medium and high effects a table of suggested values follows.

Table 1: *Cohen's d interpretation*

Effect size		Small	medium	Large
t-test for	*d*	.20	.50	.80
t-test for corr	*r*	.10	.30	.50
f-test for regress	f^2	.02	.15	.35
f-test for anova	*f*	.10	.25	.40
Chi-square	χ^2	.10	.30	.50

These values though are not suggested to be taken as absolutes and are to be interpreted within the context of the research program.

5. CONCLUSION AND RECOMENDATIONS

A good research is more than just obtaining statistical significance. To obtain the practical significance of an effect is of utmost importance. Improved research practice therefore should be to estimate effect sizes in combination with NHST techniques which helps in diminishing prevalent misuse and misinterpretation of NHST. In addition, effect size and its confidence interval provide the quantitative estimate of an effect of interest and the precision of that estimate. At the end such a process enhances more conclusive evidence in research.

Reporting and interpreting statistical evidence with effect size should be given much more focus that enables to determine the strength or impact of findings, unlike NHST that provides meaningless conclusions. Also, as scholarly journals do not accept research articles without effect size reported in them, this pushes researchers to further transcend effect size as their mandatory statistical tool in research. As a result of all, researchers in the future will and are recommended to seek not only if a sample result is likely but also if an effect is practically significant.

CONFLICT OF INTEREST

There is no conflict of interest to disclose.

ACKNOWLEDGEMENT

The authors of all papers and the books which are cited are all indebted. In addition, sincere thanks also go to all authors of any books, papers, and notes, videos etc. which were helpful during preparation of this review article.

REFERENCES

Becker, L. A. (2000). Effect size (ES) [PDF File].

Retrieved from https://www.uv.es/~friasnav/EffectSizeBecker.pdfAccessed on 19 September 2019

Danial. (2017 June 28). What Statistical Significance Really Means (10-6) [Video]. *Research by Design*. https://www.youtube.com/watch?v=wuB7CC9DrIE&t=20s

Effect Size. (2020). *Statistics Solutions*

Retrieved from https://www.statisticssolutions.com/statistical-analyses-effect-size/ March 7, 2020

Effect Size Calculator for T-Test. (2020). *Social Science Statistics. Retrieved from*

https://academic.oup.com/jpepsy/article/34/9/917/939415 *Accessed on June 25, 2020*

Effect size: What is it and when and how should I use it. (n.d.). *Physport*. Retrieved

from https://www.physport.org/recommendations/Entry.cfm?ID=93385 Accessed on March 7, 2020

Field, A. (2020 October 01). Null Hypothesis Significance Testing [Video]. *Andy Field.*

https://www.youtube.com/watch?v=IXYDMMBisr8

Field, A. (2020 October 02). Effect Sizes and Bays Factors [Video]. *Andy Field.*

https://www.youtube.com/watch?v=DjpxTmiombE&t=2051s

Grace-Martin, K. (2020). Two Types of Effect Size Statistic: Standardized and Unstandardized. *The Analysis Factor. Retrieved from* https://www.theanalysisfactor.com/two-types-effect-size-statistic/ Accessed on June 25, 2020

How is the effect size used in power analysis. (2020). *UCLA: Statistical Consulting Group*.

Retrieved from https://stats.idre.ucla.edu/other/mult-pkg/faq/general/effect-size-power/faqhow-is-effect-size-used-in-power-analysis/ (accessed March 07, 2020).

Humberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*(2), 227-240. https://doi.org/10.1177/0013164402062002002

Hypothesis Testing: Methodology and Limitations. (2001). *Elsevier Science Ltd. Retrieved from* https://www.stats.ox.ac.uk/~snijders/Encycl_isb203057.pdfAccessed on 15 December 2019

Hypothesistesting.(2011). *WikiVet*.Retrieved from https://en.wikivet.net/Hypothesis_testing#Limitations_of_null_hypothesis_tests Accessed on 15 December 2019

Durlak, J.A. (2009). How to Select, Calculate, and Interpret Effect Sizes. *Journal of Pediatric Psychology, 34* (9), 917–928, https://doi.org/10.1093/jpepsy/jsp004

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4,* Article 863.

https://doi.org/10.3389/fpsyg.2013.00863

Levine, T.R., Weber, R., Hullett, C., Park, H.S., & Massi Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34* (2), 171–187. https://doi.org/10.1111/j.1468-2958.2008.00317.x

LeCroy, C.W., & Krysik, J. (2007). Understanding and interpreting effect size measures.

Social Work Research, 31(4), 243–248, DOI: https://doi.org/10.1093/swr/31.4.243

Limitations of significance testing. (2015). *24 X 7editing.com.* Retrieved from https://www.24x7editing.com/limitations-of-the-tests-of-hypotheses/Accessed on December 13, 2019

Nickerson, R.S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods, 5*(2)*, 241-301.* DOI:10.1037/1082-989X.5.2.241

Null hypothesis significance testing. (n.d.). *InfluentialPoints.com. Biology, images, analysis, design*. Retrieved from https://influentialpoints.com/Training/null_hypothesis_significance_testing-principles-properties-assumptions.htm Accessed on 15 December 2019

Pernet, C. (2017). Null hypothesis significance testing: a short tutorial. [Abstract from FlOOOResearch]. *F1000Research, 4*. Article 621. DOI: 10.12688/f1000research.6963.5

Statistics HowTo. (2020). Effect Size (Measures of Association) Definition and Use in Research. Retrieved from https://www.statisticshowto.com/effect-size/ Accessed on 07 April 2020

Stukas, A. A., & Cumming, G. (2014). Interpreting effect sizes: Toward a quantitative cumulative social psychology. *European Journal of Social Psychology, 44*(7), pp. 711-722. doi: 10.1002/ejsp.2019

Sullivan, G. M., and Feinn, R. (2012). Using effect Size—or why the *p-value* is not enough. *Journal of Graduate Medical Education*, *4*(3), 279-282. DOI: http://dx.doi.org/10.4300/JGME-D-12-00156.1

Szucs, D., and Ioannidis, J.P.A. (2017). When null hypothesis significance testing is unsuitable for research: A reassessment. [Abstract from frontiers in Human Neuroscience]. *Frontiers in Human Neuroscience, 11.*DOI: 10.3389/fnhum.2017.00390

Téllez, A., García, C.H., and Corral-Verdugo, V. (2015). Effect size, confidence intervals and statistical power in psychological research. *Psychology in Russia: State of the Art 8*(3), pp.27 - 46. doi:10.11621/pir.2015.0303

Wasserstein, R. L., and Lazar, N.A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133. DOI: 10.1080/00031305.2016.1154108