Using Effect Size—or Why the *P* Value Is Not Enough

Gail M. Sullivan, MD, MPH
Richard Feinn, PhD

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass[1]

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen[2]

These statements about the importance of effect sizes were made by two of the most influential statistician-researchers of the past half-century. Yet many submissions to *Journal of Graduate Medical Education* omit mention of the effect size in quantitative studies while prominently displaying the *P* value. In this paper, we target readers with little or no statistical background in order to encourage you to improve your comprehension of the relevance of effect size for planning, analyzing, reporting, and understanding education research studies.

What Is Effect Size?

In medical education research studies that compare different educational interventions, effect size is the *magnitude of the difference between groups*. The *absolute effect size* is the difference between the average, or mean, outcomes in two different intervention groups. For example, if an educational intervention resulted in the improvement of subjects' examination scores by an average total of 15 of 50 questions as compared to that of another intervention, the absolute effect size is 15 questions or 3 grade levels (30%) better on the examination. Absolute effect size does not take into account the variability in scores, in that not every subject achieved the average outcome.

In another example, residents' self-assessed confidence in performing a procedure improved an average of 0.4 point on a Likert-type scale ranging from 1 to 5, after simulation training. While the absolute effect size in the first example appears clear, the effect size in the second example is less apparent. Is a 0.4 change a lot or trivial? Accounting for variability in the measured improvement may aid in interpreting the magnitude of the change in the second example.

Thus, effect size can refer to the raw difference between group means, or absolute effect size, as well as standardized measures of effect, which are calculated to transform the effect to an easily understood scale. Absolute effect size is useful when the variables under study have intrinsic meaning (eg, number of hours of sleep). Calculated indices of effect size are useful when the measurements have no intrinsic meaning, such as numbers on a Likert scale; when studies have used different scales so no direct comparison is possible; or when effect size is examined in the context of variability in the population under study.

Calculated effect sizes can also quantitatively compare results from different studies and thus are commonly used in meta-analyses.

Why Report Effect Sizes?

The effect size is the main finding of a quantitative study. While a *P* value can inform the reader whether an effect exists, the *P* value will not reveal the size of the effect. In reporting and interpreting studies, both the substantive significance (effect size) and statistical significance (*P* value) are essential results to be reported.

For this reason, effect sizes should be reported in a paper's Abstract and Results sections. In fact, 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, or β, error, which is the probability of concluding there is no effect when one actually exists. In other words, you must determine what number of subjects in the study will be sufficient to ensure (to a particular 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.

Why Isn't the *P* Value Enough?

Statistical significance is the probability that the observed difference between two groups is due to chance. If the *P* value is larger than the alpha level chosen (eg, .05), any observed difference is assumed to be explained by sampling variability. With a sufficiently large sample, a statistical test

Gail M Sullivan, MD, MPH, is Editor-in-Chief, *Journal of Graduate Medical Education;* **Richard Feinn, PhD,** is Assistant Professor, Department Psychiatry, University of Connecticut Health Center.

Corresponding author: Gail M. Sullivan, MD, MPH, University of Connecticut, 253 Farmington Avenue, Farmington, CT 06030-5215, gsullivan@nso1.uchc.edu

DOI: http://dx.doi.org/10.4300/JGME-D-12-00156.1

TABLE 1 **Common Effect Size Indices**[a]

Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's d[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (M); s is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR)	Group 1 odds of outcome / Group 2 odds of outcome If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR)	Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups	Small 2 Medium 3 Large 4	Compares probabilities of outcome occurring from one intervention to another
Measures of association			
Pearson's r correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables
r^2 coefficient of determination	Range, 0 to 1; Usually expressed as percent	Small 0.04 Medium 0.25 Large 0.64	Proportion of variance in one variable explained by the other

[a] Adapted from Ferguson et al.[9]
[b] Based on Soper.[7]

will almost always demonstrate a significant difference, unless there is no effect whatsoever, that is, when the effect size is exactly zero; yet very small differences, even if significant, are often meaningless. Thus, reporting only the significant P value for an analysis is not adequate for readers to fully understand the results.

For example, if a sample size is 10 000, a significant P value is likely to be found even when the difference in outcomes between groups is negligible and may not justify an expensive or time-consuming intervention over another. The level of significance by itself does not predict effect size. Unlike significance tests, effect size is independent of sample size. Statistical significance, on the other hand, depends upon both sample size and effect size. For this reason, P values are considered to be confounded because of their dependence on sample size. Sometimes a statistically significant result means only that a huge sample size was used.[3]

A commonly cited example of this problem is the Physicians Health Study of aspirin to prevent myocardial infarction (MI).[4] In more than 22 000 subjects over an average of 5 years, aspirin was associated with a reduction in MI (although not in overall cardiovascular mortality) that was highly statistically significant: $P < .00001$. The study was terminated early due to the conclusive evidence, and aspirin was recommended for general prevention. However, the effect size was very small: a risk difference of 0.77% with $r^2 = .001$—an extremely small effect size. As a result of that study, many people were advised to take aspirin who would not experience benefit yet were also at risk for adverse effects. Further studies found even smaller effects, and the recommendation to use aspirin has since been modified.

How to Calculate Effect Size

Depending upon the type of comparisons under study, effect size is estimated with different indices. The indices fall into two main study categories, those looking at effect sizes between groups and those looking at measures of association between variables (TABLE 1). For two independent groups, effect size can be measured by the standardized difference between two means, or mean (group 1) – mean (group 2) / standard deviation.

The denominator standardizes the difference by transforming the absolute difference into standard deviation units. Cohen's term d is an example of this type of effect size index. Cohen classified effect sizes as *small* ($d = 0.2$), *medium* ($d = 0.5$), and *large* ($d \geq 0.8$).[5] According to Cohen, "a medium effect of .5 is visible to the naked eye of

TABLE 2 **Differences between Groups, Effect Size measured by Glass's Δ[a]**

Relative Size	Effect Size	Percentile	% of Non-overlap
	0	50	0
Small	0.2	58	15
Medium	0.5	69	33
Large	0.8	79	47
	1.0	84	55
	1.5	93	71
	2.0	97	81

[a] Adapted from Bartolucci et al[4] and Coe.[6]

a careful observer. A small effect of .2 is noticeably smaller than medium but not so small as to be trivial. A large effect of .8 is the same distance above the medium as small is below it."[6] These designations large, medium, and small do not take into account other variables such as the accuracy of the assessment instrument and the diversity of the study population. However these ballpark categories provide a general guide that should also be informed by context.

Between group means, the effect size can also be understood as the average percentile distribution of group 1 vs. that of group 2 or the amount of overlap between the distributions of interventions 1 and 2 for the two groups under comparison. For an effect size of 0, the mean of group 2 is at the 50th percentile of group 1, and the distributions overlap completely (100%)—that is , there is no difference. For an effect size of 0.8, the mean of group 2 is at the 79th percentile of group 1; thus, someone from group 2 with an average score (ie, mean) would have a higher score than 79% of the people from group 1. The distributions overlap by only 53% or a non-overlap of 47% in this situation (TABLE 2).[5,6]

What Is Statistical Power and Why Do I Need It?

Statistical power is the probability that your study will find a statistically significant difference between interventions when an actual difference does exist. If statistical power is high, the likelihood of deciding there is an effect, when one does exist, is high. Power is 1-β, where β is the probability of wrongly concluding there is no effect when one actually exists. This type of error is termed Type II error. Like statistical significance, statistical power depends upon effect size and sample size. If the effect size of the intervention is large, it is possible to detect such an effect in smaller sample numbers, whereas a smaller effect size would require larger sample sizes. Huge sample sizes may detect differences that are quite small and possibly trivial.

Methods to increase the power of your study include using more potent interventions that have bigger effects, increasing the size of the sample/subjects, reducing measurement error (use highly valid outcome measures), and raising the α level but only if making a Type I error is highly unlikely.

BOX **Calculation of Sample Size Example**

Your pilot study analyzed with a Student *t*-test reveals that group 1 (N = 29) has a mean score of 30.1 (SD, 2.8) and that group 2 (N = 30) has a mean score of 28.5 (SD, 3.5). The calculated *P* value = .06, and on the surface, the difference appears not significantly different. However, the calculated effect size is 0.5, which is considered "medium" according to Cohen. In order to test your hypothesis and determine if this finding is real or due to chance (ie, to find a *significant* difference), with an effect size of 0.5 and *P* of <.05, the power will be too low unless you expand the sample size to approximately N = 60 in each group, in which case, power will reach .80. For smaller effect sizes, to avoid a Type II error, you would need to further increase the sample size. Online resources are available to help with these calculations.

How To Calculate Sample Size?

Before starting your study, calculate the power of your study with an estimated effect size; if power is too low, you may need more subjects in the study. How can you estimate an effect size before carrying out the study and finding the differences in outcomes? For the purpose of calculating a reasonable sample size, 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. There are many online sample size/power calculators available, with explanations of their use (BOX).[7,8]

Power must be calculated prior to starting the study; post-hoc calculations, sometimes reported when prior calculations are omitted, have limited value due to the incorrect assumption that the sample effect size represents the population effect size.

Of interest, a β error of 0.2 was chosen by Cohen, who postulated that an α error was more serious than a β error. Therefore, he estimated the β error at 4 times the α: $4 \times 0.05 = 0.20$. Although arbitrary, as this has been copied by researchers for decades, use of other levels will need to be explained.

Summary

Effect size helps readers understand the magnitude of differences found, whereas statistical significance examines whether the findings are likely to be due to chance. Both are essential for readers to understand the full impact of your work. Report both in the Abstract and Results sections.

References

1 Kline RB. (2004). *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington DC: American Psychological Association. p 95.
2 Cohen J. Things I have learned (so far). *Am Psychol*. 1990;45:1304–1312.
3 Ellis PD. Thresholds for interpreting effect sizes. http://www.polyu.edu.hk/mm/effectsizefaqs/thresholds_for_interpreting_effect_sizes2.html. Accessed April 16, 2012.
4 Bartolucci AA, Tendera M, Howard G. Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. *Am J Cardiol*. 2011;107(12):1796–801.
5 Carson C. The effective use of effect size indices in institutional research. http://www.keene.edu/ir/effect_size.pdf. Accessed April 16, 2012.
6 Coe R. It's the effect size, stupid: what "effect size" is and why it is important. Paper presented at the 2002 Annual Conference of the British Educational Research Association, University of Exeter, Exeter, Devon, England, September 12–14, 2002. http://www.leeds.ac.uk/educol/documents/00002182.htm. Accessed March 23, 2012.
7 Ellis PD. Effect size calculators (2009). http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html. Accessed April 15, 2012.
8 Soper D. Statistics Calculators version 3.0 beta. http://danielsoper.com/statcalc3/default.aspx. Accessed April 16, 2012.
9 Ferguson CJ. An effect size primer: a guide for clinicians and researchers. http://www.tamiu.edu/~cferguson/Ferguson%20PPRP.pdf. Accessed July 12, 2012.