Research Rundowns >Quantitative Methods > Effect Size

As you read educational research, you'll encounter *t*-test (*t*) and ANOVA (*F*) statistics frequently. Hopefully, you understand the basics of (statistical) significance testing as related to the null hypothesis and *p* values, to help you interpret results. If not, see the **Significance Testing** (*t*-test, chi square, ANOVA) review for more information. In this class, we'll consider the difference between *statistical* significance and *practical* significance, using a concept called effect size.

The "Significance" Issue

Most statistical measures used in educational research rely on some type of statistical significance measure to authenticate results. Recall that the one thing *t*-tests, ANOVA, chi square, and even correlations have in common is that interpretation relies on a *p* value (*p* = statistical significance). That is why the easy way to interpret significance studies is to look at the direction of the sign (<, =, or >) to understand if the results are statistically meaningful.

While most published statistical reports include information on significance, such measures can cause problems for practical interpretation. For example, a significance test does not tell the size of a difference between two measures (practical significance), nor can it easily be compared across studies. To account for this, the American Psychological Association (APA) recommended all published statistical reports also include effect size (for example, see the APA 5th edition manual section, 1.10: Results section). Further guidance is summed by Neill (2008):

1. *When there is no interest in generalizing (e.g., we are only interested in the results for the sample), there is no need for significance testing. In these situations, effect sizes are sufficient and suitable.*
2. *When examining effects using small sample sizes, significance testing can be misleading. Contrary to popular opinion, statistical significance is not a direct indicator of size of effect, but rather it is a function of sample size, effect size, and p level.*
3. *When examining effects using large samples, significant testing can be misleading because even small or trivial effects are likely to produce statistically significant results.*

What is Effect Size?

The simple definition of effect size is the magnitude, or size, of an effect. Statistical significance (*e.g.*, $p < .05$) tells us there was a difference between two groups or more based on some treatment or sorting variable. For example, using a *t*-test, we could evaluate whether the discussion or lecture method is better for teaching reading to 7th graders:

For six weeks, we use the discussion method to teach reading to Class A, while using the lecture method to teach reading to Class B. At the end of the six weeks, both groups take the same test. The discussion group (Class A), averages 92, while the lecture group (Class B) averages 84.

Recalling the **Significance Testing** review, we would calculate standard deviation and evaluate the results using a *t*-test. The results give us a value for *p*, telling us (if $p < .05$, for example) the discussion method is superior for teaching reading to 7th graders. What this fails to tell us is the magnitude of the difference. In other words, *how much more effective* was the discussion method? To answer this question, we standardize the difference and compare it to 0.

Effect Size (Cohen's *d, r*) & Standard Deviation

Effect size is a standard measure that can be calculated from any number of statistical outputs. One type of effect size, the standardized mean effect, expresses the mean difference between two groups in standard deviation units. Typically, you'll see this reported as Cohen's *d*, or simply referred to as *"d*." Though the values calculated for effect size are generally low, they share the same range as standard deviation (-3.0 to 3.0), so can be quite large. Interpretation depends on the research question. The meaning of effect size varies by context, but the standard interpretation offered by Cohen (1988) is:

.8 = large (8/10 of a standard deviation unit)
.5 = moderate (1/2 of a standard deviation)
.2 = small (1/5 of a standard deviation)

*Recall from the **Correlation** review r can be interpreted as an effect size using the same guidelines. If you are comparing groups, you don't need to calculate Cohen's *d*. If you are asked for effect size, it is *r*.

Calculating Effect Size (Cohen's *d*)

Option 1 (on your own)

Given mean (*m*) and standard deviation (*sd*), you can calculate effect size (*d*). The formula is:

$$d = \frac{\textbf{\textit{m1}} \text{ (group or treatment 1)} - \textbf{\textit{m2}} \text{ (group or treatment 2)}}{\textbf{\textit{[pooled] sd}}}$$

Where pooled ***sd*** is *√*sd1+sd2/*2]

Option 2 (using an online calculator)

If you have mean and standard deviation already, or the results from a *t*-test, you can use an online calculator, such as this one. When using the calculator, be sure to **only use Cohen's *d*** when you are comparing groups. If you are working with correlations, you don't need **d**. Report and interpret ***r***.

Wording Results

The basic format for group comparison is to provide: population (*N*), mean (*M*) and standard deviation (*SD*) for both samples, the statistical value (*t* or *F*), degrees freedom (*df*), significance (*p*), and confidence interval (CI.95). Follow this information with a sentence about effect size (see red, below).

Effect size example 1 (using a t-test): p ≤ .05, or *Significant Results*

Among 7th graders in Lowndes County Schools taking the CRCT reading exam (*N* = 336), there was a statistically significant difference between the two teaching teams, team 1 (*M* = 818.92, *SD* = 16.11) and team 2 (*M* = 828.28, *SD* = 14.09), *t*(98) = 3.09, *p* ≤ .05, *CI.95* -15.37, -3.35. Therefore, we reject the null hypothesis that there is no difference in reading scores between teaching teams 1 and 2. Further, Cohen's effect size value (*d* = .62) suggested a moderate to high practical significance.

Effect size example 2 (using a t-test): p ≥ .05, or *Not Significant Results*

Among 7th graders in Lowndes County Schools taking the CRCT science exam (*N* = 336), there was no statistically significant difference between female students (M = 834.00, SD = 32.81) and male students (841.08, SD = 28.76), *t*(98) = 1.15 p ≥ .05, CI.95 -19.32, 5.16. Therefore, we fail to reject the null hypothesis that there is no difference in science scores between females and males. Further, Cohen's effect size value (*d* = .09) suggested low practical significance.