Memo

To: Sharon Pearcy and other colleagues
From: Roger Bakeman
Date: March 10, 2021
Re: Cohen on small, medium, and large effects sizes

Perhaps the first thing to know about designating specific values to represent small, medium, and large effect sizes is that, although many may do it, everyone should be uneasy doing it. Cohen (1st ed. 1969, revised ed. 1977, 2nd ed., 1988) was the first to say that the values he provided are a last resort. For example, Thompson (2007), echoing Cohen, argued that it is preferable to relate effect sizes to similar effects in the literature, that different areas of research should develop their own standards for what is a meaningful effect size, but that absent such consensus (rarely seen), Cohen's guidelines remain useful—even recognizing the arbitrary way cut-points segment a continuum.

Cohen provided small, medium, and large values for several common statistics. These values are best understood as **thresholds**. An effect size is characterized as small, for example, if it is equal to or greater than the value Cohen gives for small but less than the value Cohen gives for medium. Here is Cohen's cautionary statement:

> For each statistical test's ES [effect size], the author proposes, *as a convention*, ES values to serve as operational definitions of the qualitative adjectives "small," "medium," and "large." This is an operation fraught with many dangers: the definitions are arbitrary, such qualitative concepts as "large" are sometimes misunderstood as absolute, sometimes as relative; and thus they run a risk of being misunderstood. ... Although arbitrary, the proposed conventions will be found to be reasonable by reasonable people. An effort was made in selecting these operational criteria to use levels of ES which accord with a subjective average of effect sizes such as are encountered in behavioral science. (pp. 12–13; all page numbers in this memo are for the 2nd ed., 1988)

Throughout Cohen speaks of populations and their parameters, not samples or groups, and largely leaves it for others to work out how estimates for population parameters are computed. He also routinely assumes equally numerous populations (i.e., sample sizes), although noting the robustness of statistics based on nonequal sample sizes. In this memo, I focus on four chapters: Chapters 2 and 8 (tests between means of two groups and their generalization to more than two groups and factorial designs), and Chapters 3 and 9 (correlation between two samples and its generalization to multiple regression).

Chapter 2: The *t* Test for Means

When testing the difference between means for two independent populations, Cohen's effect size (ES) statistic is **d** (and $\mathbf{d}_z$ for related populations).:

$$\mathbf{d} = (\mathbf{m}_A - \mathbf{m}_B) / \sigma \text{ and } \mathbf{d}_Z = \mathbf{m}_Z / \sigma_Z \quad \text{(Equation 2.2.1, p. 20, and Equation 2.3.5, p. 48)}$$

where $\mathbf{m}_A - \mathbf{m}_B$ is the difference between the means for the independent groups and $\mathbf{m}_Z$ is the mean of the differences between the paired samples, divided by the appropriate standard

deviation. (When quoting Cohen, I preserve his notational convention, which is to **bold** and not italicize Roman letters that represent statistics; for consistency, I do the same in my own prose.)

For Cohen's* d*, thresholds for small, medium, and large effects are 0.20, 0.50, and 0.80.

Before introducing thresholds, Cohen cautioned:

> ... there is a certain risk inherent in offering conventional operational definitions for these terms ... This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available. (p. 25).

Cohen offered a way to visualize what different values of **d** mean. Assuming two populations of normally distributed scores that are equally numerous and equally variable, he computed, the percent of nonoverlap in the populations for various values of **d**. For example, when ***d*** = .0, the nonoverlap ($\mathbf{U}_1$ in Cohen's notation) is 0%, and as **d** increases, the nonoverlap percentage increases.

But why these specific thresholds? When analyzing whether the means of two independent groups differ, a **t** test and point biserial correlation ($\mathbf{r}_p$, a correlation between a binary and a continuous variable) produce identical **p** values; they are essentially the same test. Moreover, as Cohen showed, values of **d** can be converted to **r** (equation 2.2.7, p. 24). Thus, in support of his thresholds, Cohen wrote:

> SMALL EFFECT SIZE: **d** = .2. ... When **d** = .2, normally distributed populations of equal size and variability have only 14.7% of their combined area which is not overlapped. ... From the point of view of correlation ... **d** = .2 means that the (point biserial) **r** between population membership (A vs. B) and the dependent variable **Y** is .100, and $\mathbf{r}^2$ is accordingly .010. ... The latter can be interpreted as meaning that population membership accounts for 1% of the variance in **Y** ... The above sounds indeed small ... Yet it is the order of magnitude ... [Cohen goes on to cite several behavioral science studies.]. ... (p. 25)
>
> MEDIUM EFFECT SIZE: **d** = .5. A medium effect size is conceived as one large enough to be visible to the naked eye. ... A **d** = .5 indicates that 33.0% (= $\mathbf{U}_1$) of the combined area covered by two normal equal-sized equally varying populations is not overlapped ... In terms of correlation ... **d** = .5 means that the (point biserial) **r** ... is .243. Thus, .059 (= $\mathbf{r}_2$) of the **Y** variance is "accounted for " by population membership. ... Expressed in the above terms, the reader may feel that the effect size designated medium is too small. That is, an amount not quite equal to 6% of variance may well not seem large enough to be called medium. [Again, Cohen goes on to cite several behavioral science studies.] ... (p. 26)
>
> Large EFFECT SIZE: **d** = .8. When our two populations are so separate as to make **d** = .8, almost half ($\mathbf{U}_1$ = 47.4%) of their areas are not overlapped. ... The point biserial **r** here equals .371, and $\mathbf{r}^2$ thus equals .138. Behavioral scientists who work with correlation coefficients ... do not ordinarily consider an **r** of .371 as large. Nor, in that frame of reference, does the writer. Note, however, that it is the .8 separation between means which is being designated as large, not the implied point biserial **r**. [Again, examples follow.] ... (p. 26)

I have quoted Cohen at some length here for two reasons: first, to give a sense of his reasoning in selecting particular values for small, medium, and large; and second, to note the correlation coefficient ($\mathbf{r}_p$) equivalences for his effect size statistic, which here is ***d***.

Chapter 3: The Significance of a Product Moment $\mathbf{r}_s$

When testing whether two variables are uncorrelated, the ES statistic is **r**, the Pearson product-moment correlation. (In Cohen's notation, $\mathbf{r}_s$ is the correlation coefficient obtained from a sample of **n** pairs of scores, whereas **r** is the population parameter.)

For* r*, Cohen's thresholds for small, medium, and large effects are .10, .30, and .50 absolute.

Typically, Cohen prefaces his values for small, medium, and large effects with the statement that these "are definitions for use when no others suggest themselves, or as conventions" (p. 79). Cohen then wrote:

> SMALL EFFECT SIZE: **r** = .10. An **r** of .10 in a population is indeed small. The implied **PV** [proportion of variance] is $\mathbf{r}^2$ = .01, and there seems little question but that relationships of that order in **X**, **Y** pairs in a population would not be perceptible on the basis of casual observation. ... it is comparable to the definition of a small ES for a mean difference .... [again, examples] ... (p. 79)
>
> MEDIUM EFFECT SIZE: **r** = .30. When **r** = .30, $\mathbf{r}^2$ = PV = .09, so that our definition of a medium effect in linear correlation implies that 9% of the variance in the dependent variable is attributable to the independent variable. It is shown later that this level of ES is comparable to that of medium ES in differences between two means ... (p. 80)
>
> LARGE EFFECT SIZE: **r** = .50. The definition of a large correlational ES as **r** = .50 leads to $\mathbf{r}^2$ = .25 of the variance in either variable being associated linearly with variance in the other. Its comparability with the definition of large ES in mean differences (**d** = .8) will be demonstrated below. ... (p. 80)

At first glance, the point biserial equivalences for small, medium, and large **d**—which are $\boldsymbol{r}_p$ = .100, .243, and .372—don't seem comparable to those given for **r**—.10, .30, and .50—at least for medium and large. The explanation is that the **point biserial correlation**—which assumes an underlying true dichotomy for the X or independent variable, like male–female—is not the appropriate comparison. Comparable to **r** is the **biserial correlation**—$\mathbf{r}_b$, which assumes an underlying continuous variable that has been dichotomized, like pass–fail on a test—and which is about 25% larger than $\mathbf{r}_p$ when groups are equal (i.e., when **p** = **q** = .5, at which point the height of the standard unit normal curve is .399). Specifically, $\mathbf{r}_b = 1.253\ \mathbf{r}_p$ (the square root of **pq** divided by .399) × $\mathbf{r}_p$; see Cohen & Cohen, 1983, pp. 37–39, 66–67, 521). Here is Cohen's table showing the equivalences (p. 82):

ES	**d**	$\mathbf{r}_p$	$\mathbf{r}_b$	**r**
Small	.20	.100	.125	.10
Medium	.50	.243	.304	.30
Large	.80	.371	.465	.50

As Cohen wrote, "Comparing the $\mathbf{r}_b$ equivalent to the **r** criteria of the present chapter, we find what are judged to be reasonably close values for small and large ES and almost exact equality at the very important medium ES level" (p. 82). Note, however, this equivalence assumes equally sized samples; as **p** deviates from .5, values for $\mathbf{r}_p$ and $\mathbf{r}_b$ for a specific value of **d** become less. (The sceptic in me thinks this is a cute intellectual trick.)

Chapter 8: The Analysis of Variance

Cohen's effect size index for analysis of variance is **f**, which he intends as a generalization of **d** to analysis of variance designs (i.e., designs with more than two groups or more than one factor). He did not treat repeated-measures designs explicitly (as he did in Chapter 2 with $\mathbf{d}_z$ for paired samples)—which is what makes Olejnik & Algina (2003) such a useful contribution.

Cohen defined his ES index for analysis of variance as:

$$\mathbf{f} = \frac{\sigma_{\mathbf{m}}}{\sigma}$$

that is, the ratio of the standard deviation of the population means to the overall population mean, where each population is associated with a cell in the analysis of variance design.

Cohen then defines small, medium, and large effects as **f** = .10, .25, and .40, as usual prefacing it with the caution that these thresholds are for use "When experience with a given research area or variable is insufficient to formulate alternative hypotheses as 'strong' as these procedures demand" (p. 285). When **k** = 2, **f** = ½ **d** ("i.e., the standard deviation of two values is simply half their difference," p. 276). Thus, again when **k** = 2, these thresholds are the same as **d** in Chapter 2 (one-half of .2 = .10, of .5 = .25, of .8 = .40).

The effect size **f** is uniquely Cohen's own; to my knowledge, textbooks rarely discuss it and statistical programs do not compute it. Eta squared is more commonly encountered (e.g., SPSS) as an effect size for analysis of variance results. Cohen defined η^2 in the usual way:

$\eta^2 = \frac{{\sigma_m}^2}{\sigma^2 + {\sigma_m}^2}$, which I would write as $\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}} + SS_{\text{error}}}$

He then expressed eta squared in terms of **f** (Equation 8.2.19, p. 281):

$$\eta^2 = \frac{\mathbf{f}^2}{1 + \mathbf{f}^2}$$

Thus, in terms of η^2, his thresholds for small, medium, and large are, to four digits after the decimal point, .0099, .0588, and .1379, which are usually rounded to two digits, and so:

For η^2, Cohen's thresholds for small, medium, and large effects are .01, .06, and .14

These are the values Kirk (1996) gives, citing Cohen, although he applied them to ω^2, a statistic that is essentially identical in concept to η^2 although computed slightly differently, with the result that its values are slightly less; it is generally favored by statisticians although η^2, perhaps because it is computed by SPSS, is more commonly reported.

Note also that Cohen wrote of eta squared. Partial eta squared and, in particular, generalized eta squared were in the future (e.g., Olejnik & Algina, 2003), although it seems reasonable to apply his thresholds for eta squared to partial eta squared and generalized eta squared as well.

Chapter 9: Multiple Regression and Correlation Analysis

Reflecting the generality of multiple regression–correlation (MRC)—as expressed in his ground-breaking 1968 *Psychological Bulletin* article, "Multiple regression as a general data-analytic system"—Cohen defined his ES index for multiple regression as:

$$\mathbf{f^2} = \frac{\mathrm{PV_S}}{\mathrm{PV_E}}$$

where PV_S reflects the proportion of variance accounted for by the source and PV_E the proportion accounted for by error, recognizing that PV_E could be 1 reduced by just PV_S, reduced by PV_S and other variables thought of as precursors (think step-wise regression), or further reduced by "other" variables.

Cohen then defines small, medium, and large effects as $\mathbf{f}^2$ = .02, .15, and .35, noting that "The values for $\mathbf{f}^2$ that follow [i.e., .02, .15, and .35] are somewhat larger than strict equivalence with the operational definitions for the other tests in this book would dictate" (p. 413). Indeed. The eta squared equivalences (Equation 9.2.5, p. 411), rounded to two digits, are .02, .13, and .26—not the. 01, .06, and .14 of Chapter 8 for analysis of variance. As Cohen explains:

> The reason for somewhat higher standards for $\mathbf{f}^2$ for the operational definitions in MRC is the expectation that the number of IVs in typical applications will be several (if not many). It seems intuitively evident that, for example, if $\mathbf{f}^2$ = .10 defines a "medium" $\mathbf{r}^2$ (= .09), it is reasonable for $\mathbf{f}^2$ = .15 to define a "medium" $\mathbf{R}^2$ (or partial $\mathbf{R}^2$) of .15 when several IVs are involved" (p. 413).

Intuitively evident? Frankly, I'm not sure I am completely convinced. Taking Cohen 1968 to heart, there is no difference between analysis of variance and multiple regression—if there are multiple predictors (sets of predictors) in MRC, so also in analysis of variance. In both, they are identified with degrees of freedom. So why doesn't the sauce for the goose of Chapter 8 apply to the sauce for the gander in Chapter 9?

In any event, this explains why I wrote in my 2005 *Behavior Research and Methods* article, "Recommended effect size statistics for repeated measures designs":

> Cohen, who did not consider repeated measures designs explicitly, defined an η^2 (which is not the same as Cohen's $\mathbf{f}^2$) of .02 as small, .13 as medium, and .26 as large (Cohen, 1988, pp. 413-414); it seems appropriate to apply these same guidelines to η_G^2 as well.

I would now say that this was sloppy scholarship on my part. Perhaps still under the thrall of Cohen, 1968—a must-read when I was in graduate school—I relied on Cohen's multiple regression Chapter 9 and not his analysis of variance Chapter 8 for the statement in Bakeman (2005). I now recommend the .01, .06, .14 thresholds, keeping company with Kirk (1996).

A final comment: My focus here has been on Cohen's qualitative labels of small, medium, and large. His focus in the books cited was on power analysis; small, medium, and large conventions "when no others suggest themselves" was something of an aside.

Now, 50 years later, it is worth reflecting on the dramatic changes Cohen has caused in social science. He came of age at a time when analysis of variance and multiple regression were separate worlds, hardly on speaking terms. His 1968 *Psychological Bulletin* article changed that (and—unseen by users—computer programmers quickly realized that, with a little matrix magic, common routines could undergird the computations required by various statistical procedures; think general linear model).

A few years earlier (1962), Cohen had rattled the social sciences with his revelation that articles claiming no effects (i.e., no statistically significant ones) were too often under powered—and so a power analysis became de rigueur for all grant and dissertation proposals. But now that the $p < .05$ "cliff" (you can only talk about effects below the cliff, as journal editors and reviewers are prone to admonish), has been partially demolished (thanks in part to Cohen, 1990, among others), we can realize that power analyses only matter when statistical significance is the only criterion by which results are judged. This makes only more prescient Cohen's emphasis on the size of effects.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384. https://doi.org/10.3758/bf03192707

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin, 70*, 426–443.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Revised ed.). New York: Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*(5), 746–759.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423–432. https://doi:10.1002/pits.20234

A personal note: in the mid-1960s, when I was working for the Yale Computer Center, some people in the Psychology Department, showing me formulas in Weiner (1962; *Statistical Principles in Experimental Design*), asked me to write a program for repeated-measures designs, which were not yet included in the fledging statistical packages of the time. The program (written in FORTAN) was brought to Georgia State University from Yale by a Yale post-doc GSU had hired and installed on the GSU computer system. When I asked someone where the program had come from, I got a vague, oh-we-get-them-from-here-and-there response—unaware they were talking to the author.