Memo

To: Colleagues
From: Roger Bakeman
Date: August 15, 2020 (revised March 12, 2021, and May 21, 2021)
Re: Putting Effects Sizes First: Magnitudes for Standard Statistics

My purpose in writing this memo is, first, to provide practical guidance on writing results that emphasize the magnitude of effects as recommended by most experts and, second, to explain why this is desirable and provide some references and historical background. Although the content is not specific to a particular statistical package, I am writing with the SPSS user in mind and use SPSS examples. My intent is to simplify but not mislead; all that I write can be, and has been, made considerably more complicated by experts in the field.

Practical Matters

Introduction

Results section benefit when authors define and consistently use standard terms like small, medium, and large to describe effect sizes. Standard statistical packages like SPSS often do not include appropriate effects sizes in their output, but almost always you can compute the desired effect size statistics from other statistics that are included in the output—and this memo shows you how. Doing so puts you ahead of those of our colleagues who, for whatever reason, eschew simple computation and limit themselves to only writing about results they see in the SPSS (or other statistical program) output, generally with a focus on *p* values alone.

Small, Medium, and Large Effects

All cut points are arbitrary—including the over-emphasized $p < .05$ "cliff"—but qualitative terms are nonetheless useful. Cohen has provided thresholds for what he terms small, medium, and large effects—also characterized as weak, moderate, and strong effects. Consistent use of these terms can make results sections more readable; I recommend their use. Usually I cite the second edition of Cohen's power analysis book (1988); the first edition was 1969 and the first edition revised was 1977. (*Note.* One statistic for which qualitative terms for magnitudes does not work is Cohen's kappa; see Bakeman & Quera, 2011.)

Nonetheless, Cohen was the first to say his thresholds were a last resort, to be used "when no others suggest themselves." As Thompson (2007) wrote, echoing Cohen, 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 absent such consensus (rarely seen), Cohen's terms and benchmarks remain useful.

Unsquared and Squared Effect Size Statistics

Commonly used effect size statistics are either unsquared or squared. If unsquared, they belong to either the *d* (comparison of two means) or the *r* family (bivariate association). If squared, they index proportions of variance accounted for (e.g., the R^2 of multiple regression).

Comparing Two Groups: Cohen's *d*

When comparing two groups, I recommend Cohen's *d*—the standardized difference between the means—and not a squared index because, given two groups, the standardized difference strikes me as more intuitive than a proportion of variance index.

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

The computation of Cohen's *d* is different for independent samples and paired samples. Neither is computed by SPSS, but both can be computed from statistics produced by SPSS. I usually copy-and-paste the SPSS output into Excel and enter the formula in Excel, thus letting Excel do the computational work. (*Note.* Use leading zeros because values can exceed 1.00.)

Independent Samples

For independent samples, both the *Descriptive Statistics* and the *Independent-Samples T Test* procedures in SPSS provide the needed statistics—the *M*s, *n*s, and *SD*s for the two groups.

$$d = \frac{M_1 - M_2}{\sigma} = \frac{M_1 - M_2}{\sqrt{\frac{SS_1 + SS_2}{N-2}}} = \frac{M_1 - M_2}{\sqrt{\frac{(n_1-1)\times SD_1^2 + (n_2-1)\times SD_2^2}{n_1+n_2-2}}}$$

The first formula is conceptual—*d* is the difference between the means divided by the population standard deviation. The second and third formulas are specific to two independent samples. Other formulas are possible, but the two here work whether the *n*s for the two groups are equal or unequal. The second is simple computationally but requires the sum of squares (SS) for the two groups. The third is the one usually given because the *M*s, *n*s, and *SD*s for the two groups are almost always found in the output of statistical programs like SPSS.

A refresher: *sum of squares*, symbolized *SS*, is something of an abbreviation, a nickname if you will. More fully stated, *SS* is the deviation of each raw score from the mean, squared, and then summed. The variance, symbolized S^2, is the *SS* divided by its degrees of freedom, which is $N - 1$ for a single group. The standard deviation, symbolized *SD*, is the square root of the variance. The *SS*, the S^2, and the *SD* all gauge the amount of variability in a set of scores.

Related Samples

For related samples, the *Paired-Samples T Test* procedure in SPSS provides the needed statistics—the mean of the paired differences and its *SD*. Then:

$$d_z = \frac{M_D}{\sigma_D} = \frac{M \text{ of } paired \text{ differences}}{SD \text{ of differences}}$$

Cohen's (1977, 1988) notation for independent and related samples is d and d_z, whereas Lakens (2013) uses d_s and d_z. Using d_s when appropriate could resolve ambiguity, but unfortunately is not conventional. Creating ambiguity, authors sometimes write Cohen's *d* when d_z would be more accurate (mea culpa). Although matters can be made more complex (e.g., Cummings, 2014), I recommend using d and d_z as defined here.

Considering Correlations

This is probably the easiest.

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

These guidelines apply to correlations generally. Textbooks often give separate formulas for the ϕ (phi) coefficient (correlation between two binary variables), the point biserial coefficient (correlation between a binary and a continuous or integer variable, same as a *t* test for independent samples), and the Spearman rank-order ρ (rho) correlation (raw scores replaced with ranks). But the conceptual formula for all, including *r* (the Pearson product moment correlation), is the same: the standardized scores for each pair, multiplied, summed, and averaged (where the standardized score is the mean subtracted from the raw score, divided by the standard deviation).

Squared Indices that Account for Variance: Simple and Multiple Regression.

The thresholds for *r* can be squared and applied to the squared indexes of r^2 (one predictor) and R^2 (one outcome, multiple predictors) of simple and multiple regression.

For* R^2*, Cohen's thresholds for small, medium, and large effects are .01, .09, and .25.

These thresholds also apply to the ΔR^2 of stepwise or hierarchic multiple regression.

Squared Indices that Account for Variance: Analysis of Variance

With more than one group, whether independent or related, we could use multiple pairwise comparisons with Cohen's *d*—or we could use a squared index. The squared index is usually preferred, especially when a design includes more than one variable, because it provides a single, omnibus index for each variable and so is less piecemeal than multiple *d*s. For analysis of variance designs, a squared index is the usual choice. Squared indexes include η^2 (eta squared), ω^2 (omega squared), and their variants.

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

Which variant of eta squared you choose—eta squared, partial eta-squared, or generalized eta squared—depends on your design and the particular effect whose magnitude you want to assess. Use **eta squared** only if you want to assess the strength of the single effect in a one-way design—that is, a design with a single factor whose levels indicate independent groups (in which case $\eta^2 = R^2$ when predictors code for group membership). For more complex designs, use **partial eta squared** if you want to assess the strength of main effects and interactions that do not include repeated-measures or covariates, and use **generalized eta squared** if you want to assess the strength of main effects and interactions that do include repeated-measures, covariates, or both.

Here are the formulas:

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}}, \quad \eta_p^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}}+SS_{\text{error}}}, \quad \eta_G^2 = \frac{SS_{\text{effect}}}{SS_{\text{effect}}+SS_{\text{error}}+SS_{\text{subjects}}}$$

The only difference is the denominator and, as you might guess, when the effect contains no repeated factors, values for η_G^2 are same as for η_p^2, and for the single effect of a one-way design, $\eta_p^2 = \eta^2$. I am engaging in some simplification here. The formula for η_G^2 is usually presented in more complex form (see Olejnik & Algina, 2003; Bakeman, 2005). I am also assuming that any between-subjects variables are fixed—levels chosen rationally and not randomly—which is the usual case (see Keppel, 1991, Keppel & Wickens, 2004) and that no covariates or other measured (as opposed to manipulated) variables are involved. For additional detail, see the cited articles.

Distinctions among these variants of eta squared are not always made. When SPSS first introduced partial eta squared in its analysis of variance output, it was labeled eta squared, not partial eta squared, and it is still not unknown for authors to write eta squared when partial eta squared would be correct (mea culpa). SPSS output gives partial eta squared for analysis of variance effects without repeated measures—which is fine—but also for effects involving repeated measures—less fine—and it is not unknown for authors to give partial eta squared when generalized eta squared would be correct (mea culpa).

Reporting η_p^2 instead of η_G^2 for effects involving repeated measures (or covariates) results in an overestimate of effect size and a misapplication of Cohen's eta squared guidelines: "As Olejnik and Algina (2003) explain ... using these benchmarks [i.e., Cohen's] when interpreting the η_p^2 effect size in designs that include covariates or repeated measures is not consistent with the considerations upon which the benchmarks were based" (Lakens, 2013, p. 7).

The good news is that η_G^2, like Cohen's *d* and d_z, is easily computed from statistics provided by the *General Linear Model–Repeated Measures* procedure in SPSS: Analyze > General Linear Model > Repeated Measures. If you select *Options* and check *Estimates of effect size*, the SPSS output will give *Partial Eta Squared*.

How to Compute Generalized Eta Squared from SPSS Output

To compute generalized eta squared for an effect:
1. select the Sum of Squares for that effect from the **Tests of Within-Subjects Effects** table, then
2. divide it by the Sum of Squares for that effect plus the error for that effect, also from the **Tests of Within-Subjects Effects** table, plus the error from the **Tests of Between-Subjects Effects** table (this is the SS_{subjects} in the formula for η_G^2 given above).

If you divide the SS for the effect just by the sum of the SS for the effect plus the SS for error from the **Tests of Within-Subjects Effects** table, you will get the value for η_p^2 given in the output.

Computing Generalized Eta Squared: An Example

This example uses a measure of behavioral regulation assessed for 47 children, 34 born preterm and 14 at term, when they were 5, 6, and 7 years of age—thus a one-between, one-within

mixed design (thanks to Natacha Akshoomoff and Carolyn Sawyer for the data). Here is the SPSS output (cleaned up a bit).

Tests of Within-Subjects Effects

Source		Type I Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Generalized Eta Squared
Age	Sphericity	14348	2	7174.2	69.118	.000	.606	.393
Age * Term	Sphericity	530.53	2	265.27	2.556	.083	.054	.023
Error(Age)	Sphericity	9341.7	90	103.8				

Tests of Within-Subjects Contrasts

Source		Type I Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Generalized Eta Squared
Age	Linear	13826	1	13826	93.617	.000	.675	.415
	Quadratic	522.89	1	522.89	8.728	.005	.162	.044
Age * Term	Linear	524.8	1	524.8	3.554	.066	.073	.026
	Quadratic	5.7298	1	5.7298	.096	.759	.002	.0004
Error(Age)	Linear	6645.7	45	147.68				
	Quadratic	2696	45	59.912				

Tests of Between-Subjects Effects

Source	Type I Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	223684	1	223684	785.59	.000	.946
Term	3786.3	1	3786.3	13.298	.001	.228
Error	12813	45	284.73			

Here I have added the Generalized Eta Squared column and have highlighted the error terms. The between-subjects error is the $SS_{subjects}$ in the formula given above. You can verify that the generalized eta squared for the 2 *df* age effect is

14,348 divided by (14,348 + 9341.7 + 12,813) = .393,

that the 1 *df* linear component of the age effect is

13,826 divided by (13,826 + 6645.7 + 12813) = .415,

and that the partial eta squared for the between-subjects term effect is

3,786 divided by (3,786 + 12,813) = .228.

The linear trend was strong, the quadratic trend (in this case an inverted U) was only weak. Nonetheless, the influence of the inverted-U trend is seen in the differences between the ages. The difference between the age 5 and 6 means was strong (d_z = 1.22), whereas the difference between the age 6 and 7 means was moderate (d_z = 0.64)—which illustrates how different effect sizes can combined when explicating findings.

Omega Squared: An Alternative to Eta Squared

As experts have long noted, η^2 is biased—its values provide an overestimate. Hays (1963) proposed ω^2 as a less-biased estimate, and since then statistical writers often discuss ω^2 and

hardly mention η^2 if at all (e.g., Keppel, 1991, Keppel & Wickens, 2004, Kirk, 1982). Following Hays, the usual formula includes mean squares. Here is an equivalent, alternative formula, written to emphasize its similarity to the formulas for η^2 given earlier:

$$\omega_{\mathrm{p}}^2 = \frac{SS_{\mathrm{effect}} - SS_{\mathrm{error}} \times \frac{df_{\mathrm{effect}}}{df_{\mathrm{error}}}}{SS_{\mathrm{effect}} + SS_{\mathrm{error}} \times \frac{N - df_{\mathrm{effect}}}{df_{\mathrm{error}}}}$$

Typically, the numerator will be somewhat less and the denominator slightly more than for η_{p}^2, thus values for ω_{p}^2 are usually less than for η_{p}^2. For example, for the term by age analysis given earlier, the η_{p}^2 for the term effect was .228 but the ω_{p}^2 for the same effect was .207—still, both large effects.

If ω^2 is less biased than η^2, why has it not become the standard effect size reported, replacing η^2? I suspect it has something to do with the tendency I mentioned earlier of colleagues who limit themselves to writing only about results they see in the SPSS (or other statistical program) output. Had SPSS output included ω^2 and not η^2, we might now see ω^2 more frequently in research reports—although, it is also true that we see η^2 less frequently than we should.

The two are very similar. Both η^2 and ω^2 have partial and generalized variants and ω_{G}^2 should be used with repeated measures designs. But ω_{G}^2 is more complex to compute than η_{G}^2, which is why I only gave formulas for η_{G}^2 in my article (Bakeman, 2005). Here is Lakens (2013) advice:

> Calculating generalized omega squared and (ω_{G}^2) can become rather complex, depending on the design (see the lists of formulas provided by Olejnik and Algina,2003). Given this complexity, and the relatively small difference between the bias and less biased estimate, I recommend researchers report η_{G}^2 ... at least until generalized omega-squared is automatically provided by statistical software packages. (p. 6)

Cohen's η^2 Benchmarks

Cohen's benchmarks for *d* and *r* are clearly marked with upper-case letters and paragraph headings (Cohen, 1977, 1988). Not so for η^2. Cohen proposed another effect size index, **f**, which has hardly been used by others, and he barely mentions η^2. He does give small, medium, and large thresholds for **f**, however, and from these and an equation he provides (1988, Equation 8.2.19, p. 281), corresponding values for η^2 can be derived. Rounded to four digits they are .0099, .0588, and .1379. Rounded to two digits they are .01, .06, and .14.

Thus rounded to three digits they are .010, .059, and .138, which are the values Kirk (1996) gives, citing Cohen, but applying them to ω^2. Confusing η^2 and ω^2 is not uncommon. Lakens (2013), for example, wrote that "Keppel (1991) has recommended partial eta squared (η_{p}^2) to improve the comparability of effect sizes between studies (p. 5), but Keppel never discussed η^2, only ω^2, and gives benchmarks of .01, .06, and .15 (p. 66), again citing Cohen. Rabbit holes abound.

Why it Matters

Why emphasize effect sizes? Let me quote from a chapter, *The Practical Importance of Research Findings*, that I wrote for an SRCD monograph (Bakeman, 2006).

> Increasingly, those who reflect on research practice have come to view a narrow focus on the statistical significance of findings, a focus that has dominated much of our thinking about data analysis for more than half a century, as something of a will-o'-the-wisp, beckoning us further into the swamp of false certainty (e.g., Cohen, 1990, 1994; Rosnow, & Rosenthal, 1989; Wilkinson et al., 1999).
>
> The traditional landscape of statistical inference has been dominated by the .05 cliff and a fear of saying things that are not so. Most of us were taught to set our alpha level at the conventional .05, thereby insuring that the probability of making a false claim (a Type I error) would be .05. Not emphasized was the conditional nature of this statement; the probability is .05 if and only if there genuinely is no effect. As Cohen pungently reminded us (1990), if something is so, claiming an effect is never false; when there is an effect, however small, the probably of a Type I error is zero; and "So if the null hypothesis is always false, what's the big deal about rejecting it?" (p. 1308).
>
> The big deal is, too many of us—researchers, reviewers, and journal editors alike—appear to have embraced the false belief that if a result is not statistically significant, then there is no effect; any apparent effect is just chance. If a result is significant, $p < .05$, we can discuss it; if not, and we mention it, journal editors may slap our hands. It is as though Cohen (1990) never wrote, ".05 is not a cliff but a convenient reference point along the possibility–probability continuum"(p. 1311), or as Rosnow and Rosenthal (1989) quipped, "surely, God loves the .06 nearly as much as the .05" (p. 1277). The all-or-nothing approach—statistical significance indicates an effect but insignificance indicates no effect—permeates especially introductions to empirical reports, which too often inform us that A affects B but does not affect C, and base such all-or-nothing statements on whether or not statistical significance was reported in previous literature. However, statistical significance should not overly impress us. After all, even the most miniscule effect can achieve statistical significance if the sample size is large enough. ...
>
> The advantages [of emphasizing effect sizes] are overwhelming (Wilkinson et al., 1999). When effect sizes occupy center-stage in results sections, accumulating patterns of results across studies are more evident—and more likely to provide the needed grist for a subsequent meta-analyst's mill. The dark influence of the under-discussed variable of sample size, which leads us to regard as real a correlation of .30 only when the sample size is 44 or more, is minimized, as are other distorting influences on our literature (Schmidt, 1996). Further, when effect sizes are emphasized we are more likely to confront issues of real-world importance, and perhaps in the process argue that in some areas of research seemingly small effect sizes can have important practical consequences (McCartney & Rosenthal, 2000), which can only deepen understanding the practical importance of our findings.
>
> Why are effect sizes not more emphasized, given their advantages? McCartney and Rosenthal (2000) suggested two reasons: knowledge and sense. Investigators may lack the knowledge required to compute and report estimates of effect sizes and, beyond knowledge, may lack an intuitive sense of how to interpret the magnitude of effects once computed. Knowledge is more easily remedied. ... [My motivation for writing this memo is to increase knowledge.]

For 20 years the Publication Manual of the American Psychological Association has been offering similar advice:

> For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section" (American Psychological Association [APA], 2000, 5th ed., p. 25).
>
> For readers to appreciate the magnitude or importance of a study's findings, it is recommended to include some measure of effects size in the Results section" (American Psychological Association [APA], 2020, 7th ed., p. 89).

There are many citations I could add. A few are Cohen (1990, 1994), Kirk (1996), Lakens (2013), Rosnow and Rosenthal (1989), Schmidt (1996), and Wilkinson (1999). A particularly incisive critique of null hypothesis significance testing is offered by Cumming (2014), as Frank Haist reminded me. He also recommends Ellis (2010). All these are worth reading, but my sentimental favorite is Cohen (1990).

Concluding Comments

A good summary statement is this: "The techniques are not new, but adopting them widely would be new for many researchers, as well as highly beneficial" (Cummings, 2014, p. 7). To be sure, there are complexities and controversies in the technical literature that I have elided here, but I don't think I have said anything that would lead you astray. Many contributors to the effect size literature are motivated by improving meta analytic studies, for which effect size indexes that are comparable over different studies and designs are essential. My motivation is to encourage us to emphasize effect sizes more when writing results. Absent the ideal of consensus in our field as to what magnitudes of effects are noteworthy—and acknowledging that Cohen wrote, "The values chosen had no more reliable a basis than my own intuition ... there are difficulties and much room for misunderstanding" (Cohen, 1988, p. 532)—it makes sense to use his small, medium, and large benchmarks, for which again, effect sizes that are comparable over different studies and designs are essential. This is especially important for repeated-measures designs, which require generalized and not partial eta squared.

More Examples

Three recent papers that have used these techniques are:

Adamson, L. B., Suma, K., Bakeman, R., Kellerman, A., & Robins, D. L. (2001). Auditory joint attention skills: Development and diagnostic differences during infancy. *Infant Behavior and Development*. https://doi.org/10.1016/j.infbeh.2021.101560

Goodman, S. H., Bakeman, R., & Milgramm, A. (2021). Continuity and stability of parenting of infants by women at risk for perinatal depression. *Parenting: Science and Practice*. https://doi.org/10.1080/15295192.2021.1877991

Sawyer, C., Adrian, J., Bakeman, R., Fuller, M., & Akshoomoff, N. (2021). Self-regulation task in young school age children born preterm: Correlation with early academic achievement. *Early Human Development*. https://doi.org/10.1016/j.earlhumdev.2021.105362

Acknowledgement. Thanks to Daryl Nenstiel and Frank Haist for comments on an earlier draft.

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

Bakeman, R. (2006). The practical importance of research findings. In K. McCartney, M. R. Burchinal, & K. L. Bub (Eds.), Best Practices in Developmental Research Methods (pp. 128–146). *Monographs of the Society for Research in Child Development, 71*(3, Serial No. 285).

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

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

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist, 49, 997–1003.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1):7–29. https://doi.org/10.1177/0956797613504966.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. New York: Cambridge University Press.

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart and Winston.

Keppel, G. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englwood Cliffs, NJ: Prentice Hall.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral science.* Belmont, CA: Brooks/Cole.

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

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–11. https://doi: 10.3389/fpsyg.2013.00863

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in the psychological sciences. *American Psychologist, 44*, 1276–1284

Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129.

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

Wilkinson, L., & the Task Force on Statistical Inference, American Psychological Association Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.