

Standardized Effect Size Estimation: Why and How?®

Statisticians have long opined that researchers generally need to present estimates of the sizes of the effects which they report. It is not sufficient to report whether or not an effect is “statistically significant.” Statistical significance only tells you that the obtained sample results would be highly unlikely were the tested (aka “null”) hypothesis absolutely true. The tested hypothesis is almost always that the size of the effect in the population from which the sample data were randomly drawn is exactly zero.

The probability of finding statistical significance (power) is a function of the size of the effect in the population, the number of data points in the sample, the research design, characteristics of the statistical test being employed, and assorted other things such as to what extent the assumptions of the statistical test have been violated. When sample sizes are large and/or other factors lead to great power, it is quite possible to obtain results which are statistically significant even when the size of the effect is trivial. The word “significant” can be very misleading in this case. If you give your audience estimates of the size of the significant effects that you have obtained, they should be able to draw their own conclusions regarding whether or not the effects are nontrivial in size.

It is also possible to obtain results that are not statistically significant even though the size of the effect in the population is large enough to be of interest. In this case it is also useful to report an estimate of the size of the effect in the population, even if you have not been convinced that it could not possibly be zero. This will help you and others decide whether or not it is worth the effort to follow-up your research with additional research with larger sample sizes or a more powerful design.

Student’s T Tests

Even in complex research it is often the case that the most important questions to be answered can be addressed by a series of comparisons between pairs of means. There are several strength of effect estimates that can be employed for this purpose. My favorite is the standardized difference between group means. Unfortunately there is much confusion about the notation used to symbolize this estimate. I shall employ the following conventions:

- The population parameter is Cohen’s δ .
- The sample estimate of Cohen’s δ will be symbolized with the letter d , or, to emphasize that it is an estimate, with \hat{d} .
- Please note that there is much notational variance in the literature.

One Sample Estimation

Here you are estimating the difference between one population mean and some constant. For example, I have sample data on the SAT-Quantitative scores of students in my undergraduate statistics class. I know that the mean SAT-Q score nationwide is 516. I wish to estimate by how much μ for my students differs from the national average. My students’ sample mean exceeds 516 by 18.78 points, and the *SD* is 93.385.

$$\hat{d} = \frac{M - \mu_{\phi}}{s} = \frac{18.78}{93.385} = .20$$

The point estimate of the effect size is .2. The point estimate does not give any indication of how much error there is. To correct this deficiency, one should provide a confidence interval for the effect size. Although methods for calculating approximate confidence intervals for d have been around for a long time, it is only recently that methods for constructing exact confidence intervals have been developed. The approximate method involves finding the unstandardized confidence interval and then dividing each end of the interval by the sample SD . This may be adequate when sample sizes are large and thus the estimation of the population SD good. With smaller sample sizes one needs to estimate both the population mean and the population SD , and the exact methods do this with a computer-intensive iterative procedure.

I have made available programs for constructing the standardized confidence interval with [SPSS](#) or [SAS](#). For this example, the obtained 95% CI runs from .015 to .316.

Presenting the Results. The mean math SAT of my undergraduate statistics students ($M = 535$, $SD = 93.4$) was significantly greater than the national norm (516), $t(113) = 2.147$, $p = .034$, $d = .20$. A 95% confidence interval for the mean runs from 517 to 552. A 95% confidence interval for d runs from .015 to .386.

Benchmarks for Standardized Differences Between Means. Is the value of the effect size estimate trivial, small, medium, large, or gargantuan? That really depends on the context of the research. In some contexts a d of .20 would be considered small but not trivial, in others it would be considered very large. That said, J. Cohen did, reluctantly, provide the following benchmarks for behavioral research:

- .2 is small, but not trivial
- .5 is medium
- .8 is large.

Reducing Error in the Effect Size Estimate. Notice that the confidence interval for the standardized effect runs from .015 (trivial) to .316 (small to medium). If that makes you uncomfortable (it should), you might want to know how to make the interval less wide. The answer is to do any or all of those things that increase power, such as increasing sample size.

Why Standardize? Statisticians frequently argue about the answer to this question. Those who work in areas where the variables of interest are measured in meaningful units (such as grams, meters, dollars and Euros, etc.) think it foolish and deceptive to standardize effect size estimates. It think they are absolutely correct for the sort of data they deal with. If I tell you that my new weight reduction program produces, on average, 17.3 lbs. more weight loss than does the typical program, you have a good feel for how large an effect that is. On the other hand, if I tell you that residents of Mississippi score, on average, 17.3 points higher than the national norm on a survey instrument designed to measure neofascist attitudes, you have no idea whether that is a trivial difference or an enormous difference. It would help to standardize that effect size estimate.

Sample Homogeneity, Extraneous Variable Control, and Standardized Effect Size Estimates. Much unappreciated is the fact that the magnitude of the effect size estimate can be

greatly influenced by the extent to which the researcher has been able to eliminate from the data the effects of various extraneous variables. When the data are collected in the laboratory, with subjects who do not differ greatly from one another on extraneous variables, and where the influence of other variables is reduced by isolating the subjects from them (things like noise, time of day, etc.), the *SD* of the scores can be greatly reduced relative to what it would be out in the natural world. Since this *SD* is the denominator (standardizer) of the effect size estimate, this can result in the effect size estimate greatly overestimating what it would be in the natural world. Imagine the difference between means is 25. For data collected in the lab, the *SD* is 15 and $d = 1.67$, a whopper effect. For data collected in the field, the *SD* is 100 and $d = .25$, a small effect.

Two Sample Estimation, Independent Samples

The basics here are the same as for the one sample situation, but you have two sample means. You wish to estimate the difference between the two corresponding population means.

Estimated Cohen's δ . The parameter being estimated here is $\delta = \frac{\mu_1 - \mu_2}{\sigma}$. The estimator is

$\hat{d} = \frac{M_1 - M_2}{s_{pooled}}$, where the pooled standard deviation is the square root of the within groups mean

square (from a one-way ANOVA comparing the two groups). If you have equal sample sizes, the pooled standard deviation is $s_{pooled} = \sqrt{.5(s_1^2 + s_2^2)}$. If you have unequal sample sizes,

$s_{pooled} = \sqrt{\sum(p_j s_j^2)}$, where for each group s_j^2 is the within-group variance and $p_j = \frac{n_j}{N}$, the proportion of the total number of scores (in both groups, N) which are in that group (n_j). You can also compute

estimated d as $\hat{d} = \frac{t\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}}$, where t is the pooled variances independent samples t comparing the two group means.

You can use the program [Conf Interval-d2.sas](#) (SPSS users, download [CI-d-SPSS.zip](#)) to obtain the confidence interval for the standardized difference between means. It will require that you give the sample sizes and the values of t and df . Use the pooled variances (equal variances assumed) values of t and df . Why the pooled variances t and df ? See [Confidence Intervals, Pooled and Separate Variances T](#).

I shall illustrate using data comparing boys' GPA with girls' GPA (participants were students in Vermont). Please look at the [computer output](#). For the girls, $M = 2.824$, $SD = .83$, $n = 33$, and for the boys, $M = 2.236$, $SD = .81$, $n = 55$. On average, the girls' GPA is .588 points higher than the boys'.

$$s_{pooled} = \sqrt{\frac{33}{88}(.83)^2 + \frac{55}{88}(.81)^2} = .818.$$

$$\hat{d} = \frac{2.824 - 2.237}{.818} = .72. \text{ Also, } \hat{d} = \frac{3.267\sqrt{33+55}}{\sqrt{33(55)}} = .72. \text{ This falls just short of a large effect by}$$

Cohen's guidelines.

Using our example data, a succinct summary statement should read something like this: Among Vermont school-children, girls' GPA ($M = 2.82$, $SD = .83$, $N = 33$) was significantly higher than boys' GPA ($M = 2.24$, $SD = .81$, $N = 55$), $t(65.9) = 3.24$, $p = .002$, $d = .72$. A 95% confidence interval for the difference between girls' and boys' mean GPA runs from .23 to .95 in raw score units and from .27 to 1.16 in standardized units.

Glass' Delta. $\Delta = \frac{M_1 - M_2}{S_{control}}$. That is, in computing the standardized difference between group

means we use the control group standard deviation in the denominator. This makes good sense when you have reasons to believe that the control group SD is a better estimate of the SD in the population to which you wish to infer the results than is the experimental group SD .

Point-Biserial r . This is simply the Pearson r between the grouping variable (coded numerically) and the criterion variable. It can be computed from the pooled variances independent t :

$r_{pb} = \sqrt{\frac{t^2}{t^2 + df}}$. For the comparison between girls' and boys' GPS, $r_{pb} = \sqrt{\frac{3.267^2}{3.267^2 + 86}} = .332$. This is

the standardized slope for the regression line for predicting the criterion variable from the grouping variable. The unstandardized slope is the difference between the group means. We standardize this difference by multiplying it by the standard deviation of the grouping variable and dividing by the standard deviation of the criterion variable. For our comparison, $r_{pb} = \frac{.588(.487)}{.861} = .33$, which is a medium-sized effect by Cohen's guidelines. Hmm, large by one criterion but medium by another – more on this later.

Eta-squared. For a two-group design, this is simply the squared point-biserial correlation coefficient. It can be interpreted as the proportion of the variance in the criterion variable which is explained by the grouping variable. For our data, $\eta^2 = .11$. For a confidence interval, use my program [Conf-Interval-R2-Regr.sas](#) (SPSS users, download [CI-R2-SPSS.zip](#)). It will ask you for F (enter the square of the pooled t), df_num (enter 1), and df_den (enter the df for the pooled t). For our data, a 95% confidence interval runs from .017 to .240.

Earlier I mentioned that the magnitude of the effect size estimate can be greatly influenced by the extent to which the researcher has managed to eliminate the effects of extraneous variables. This applies to proportion of variance effect size estimates (and similar statistics) every bit as much as it does to estimates of the standardized difference between means.

Omega-squared. Eta-squared is a biased estimator, tending to overestimate the population parameter. Less biased is the omega-squared statistic.

Point Biserial r versus Estimated d . Each of these has its advocates. Regardless of which you employ, you should be aware that the ratio of the two samples sizes can have a drastic effect on the value of the point-biserial r , but does not affect the value of estimated d . See [Effect of \$n_1/n_2\$ on Estimated \$d\$ and \$r_{pb}\$](#) .

Common Language Effect Size Statistic. See <http://core.ecu.edu/psyc/wuenschk/docs30/CL.pdf>. CL is the probability that a randomly selected score from the one population will be greater than a randomly sampled score from the other

distribution. Compute $Z = \frac{|M_1 - M_2|}{\sqrt{s_1^2 + s_2^2}}$ and then find the probability of obtaining a Z less than the

computed value. For the data here, $Z = \frac{|2.82 - 2.24|}{\sqrt{.83^2 + .81^2}} = 0.50$, which yields a lower-tailed p of .69.

That is, if one boy and one girl were randomly selected, the probability that the girl would have the higher GPA is .69. If you prefer odds, the odds of the girl having the higher GPA = $.69/(1-.69) = 2.23$ to 1.

Two Sample Estimation, Correlated Samples

Treat the data as if they were from independent samples. If you base your effect size estimate on the correlated samples analysis, you will overestimate the size of the effect. You cannot use my [Conf. Interval-d2.sas](#) program to construct a confidence interval for d when the data are from correlated samples. With correlated samples the distributions here are very complex, not following the noncentral t . You can construct an approximate confidence interval, $\hat{d} \pm z_{cc} SE$ where SE is

$\sqrt{\frac{d^2}{2(n-1)} + \frac{2(1-r_{12})}{n}}$, but such a confidence interval is not very accurate unless you have rather large sample sizes.

You could compute $\hat{d}_{diff} = \frac{M_1 - M_2}{s_{diff}}$, where s_{Diff} is the standard deviation of the difference

scores, but this would artificially inflate the size of the effect, because the correlation between conditions will probably make s_{diff} smaller than the within-conditions standard deviation. The bottom line here is this: The researcher having gained greater power by using a correlated samples design does not make the actual difference between the populations means any greater than it would be if the researcher had used an independent samples design, so, IMHO, the used of s_{diff} as the standardizer cannot usually be justified.

Correlation/Regression Analysis

Even in complex research, it is often the case that the most important questions to be answered can be addressed by a series of correlations between pairs of variables. [Does that statement sound familiar? It should.] The statistic most often employed here is the Pearson r , and the corresponding parameter is Pearson ρ . Pearson r and Pearson ρ are already standardized – they represent the number of standard deviations the one variable changes for each one standard deviation change in the other.

Benchmarks for ρ . Again, context can be very important.

- .1 is small but not trivial
- .3 is medium
- .5 is large

Confidence Interval for ρ , Correlation Analysis. My colleagues and I (Wuensch, K. L., Castellow, W. A., & Moore, C. H. Effects of defendant attractiveness and type of crime on juridic

judgment. *Journal of Social Behavior and Personality*, 1991, 6, 713-724) asked mock jurors to rate the seriousness of a crime and also to recommend a sentence for a defendant who was convicted of that crime. The observed correlation between seriousness and sentence was .555, $n = 318$, $p < .001$. We treat both variables as random. Now we roll up our sleeves and prepare to do a bunch of tedious arithmetic.

First we apply Fisher's transformation to the observed value of r . I shall use Greek zeta to symbolize the transformed r .

$$\zeta = (0.5) \ln \left| \frac{1+r}{1-r} \right| = (0.5) \ln \left| \frac{1.555}{.445} \right| = (0.5) \ln(3.494) = (0.5)(1.251) = 0.626. \text{ We compute the standard}$$

error as $SE_r = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{315}} = .05634$. We compute a 95% confidence interval for zeta:

$$\zeta \pm z_{cc} SE_r = .626 \pm 1.96(.05634). \text{ This give us a confidence interval extending from .51557 to .73643}$$

– but it is in transformed units, so we need to untransform it. $r = \frac{e^{2\zeta} - 1}{e^{2\zeta} + 1}$. At the lower boundary, that

$$\text{gives us } r = \frac{e^{1.031} - 1}{e^{1.031} + 1} = \frac{1.8039}{3.8039} = .474, \text{ and at the upper boundary } r = \frac{e^{1.473} - 1}{e^{1.473} + 1} = \frac{3.3617}{5.3617} = .627.$$

What a bunch of tedious arithmetic that involved. We need a computer program to do it for us. See [Weaver and Wuensch](#) for SAS and SPSS code to do this for you.

Suppose that we obtain $r = .8$, $n = 5$. Using my SAS program, the 95% confidence interval runs from -0.28 to +0.99.

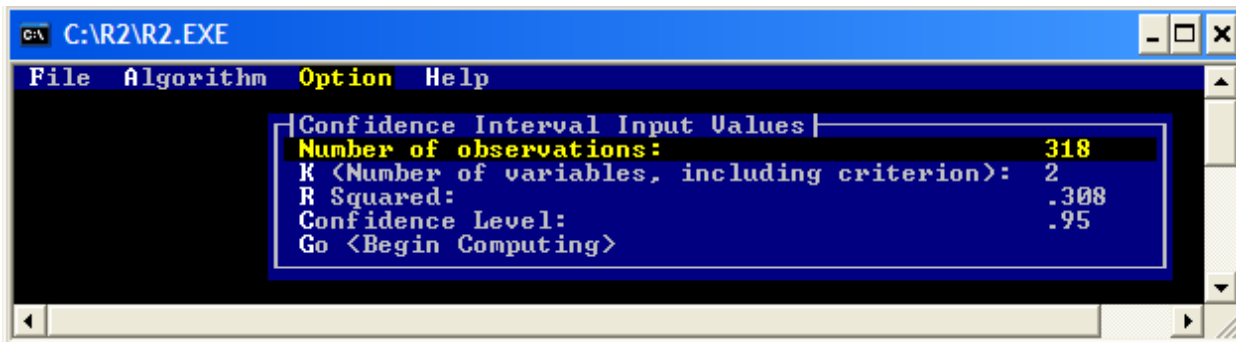
Don't have SAS or SPSS? Try [the calculator at Vassar](#).

r =	<input type="text" value=".555"/>	<input type="button" value="Reset"/> <input type="button" value="Calculate"/>
n =	<input type="text" value="318"/>	

0.95 and 0.99 Confidence Intervals of rho

	Lower Limit	Upper Limit
0.95	0.474	0.626
0.99	0.447	0.647

Confidence Interval for ρ^2 , Correlation Analysis. Use R2, which is available for free, from James H. Steiger and Rachel T. Fouladi. You can download the program and the manual [here](#). Unzip the files and put them in the directory/folder R2. Navigate to the R2 directory and run (double click) the file R2.EXE. A window will appear with R2 in white on a black background. Hit any key to continue. Enter the letter O to get the Options drop down menu. Enter the letter C to enter the confidence interval routine. Enter the letter N to bring up the sample size data entry window. Enter 318 and hit the enter key. Enter the letter K to bring up the number of variables data entry window. Enter 2 and hit the enter key. Enter the letter R to enter the R^2 data entry window. Enter .308 (that is .555 squared) and hit the enter key. Enter the letter C to bring up the confidence level data entry window. Enter .95 and hit the enter key. The window should now look like this:

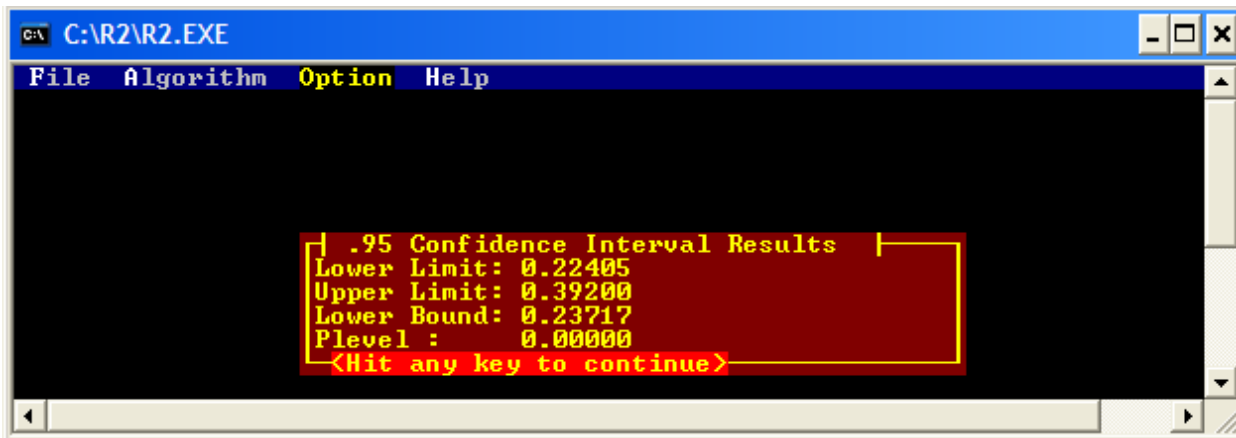


```

C:\R2\R2.EXE
File Algorithm Option Help
[Confidence Interval Input Values]
Number of observations: 318
K (Number of variables, including criterion): 2
R Squared: .308
Confidence Level: .95
Go <Begin Computing>

```

Enter G to begin computing. Hit any key to display the results.



```

C:\R2\R2.EXE
File Algorithm Option Help
|.95 Confidence Interval Results|
Lower Limit: 0.22405
Upper Limit: 0.39200
Lower Bound: 0.23717
Plevel : 0.00000
<Hit any key to continue>

```

As you can see, we get a confidence interval for r^2 that extends from .224 to .393.

Hit any key to continue, F to display the File drop-down menu, and X to exit the program.

The R2 program will not handle sample sizes greater than 5,000. In that case you can use the approximation procedure which is programmed into my SAS program [Conf-Interval-R2-Regr-LargeN.sas](#). This program assumes a regression model (fixed predictors) rather than a correlation model (random predictors), but in my experience the confidence intervals computed by R2 differ very little from those computed with my large N SAS program when sample size is large (in the thousands).

What Confidence Coefficient Should I Employ? When dealing with R^2 , if you want your confidence interval to correspond to the traditional test of significance, you should employ a confidence coefficient of $(1 - 2\alpha)$. For example, for the usual .05 criterion of statistical significance, use a 90% confidence interval, not 95%. This is illustrated below.

Suppose you obtain $r = .26$ from $n = 62$ pairs of scores. If you compute t to test the null that ρ is zero in the population, you obtain $t(60) = 2.089$. If you compute F , you obtain $F(1, 60) = 4.35$. The p value is .041. At the .04 level, the correlation is significant. When you put a 95% confidence interval about r you obtain .01, .48. Zero is not included in the confidence interval. Now let us put a confidence interval about the r^2 (.0676) using Steiger & Fouladi's R2.

```

|Confidence Interval Input Values|
|Number of observations:          62
|K (Number of variables, including criterion): 2
|R Squared:                      .0676
|Confidence Level:              .95
|Go <Begin Computing>

```

```

|.95 Confidence Interval Results
|Lower Limit: 0.00000
|Upper Limit: 0.22612
|Lower Bound: 0.00145
|Plevel :    0.04127

```

Oh my, the confidence interval includes zero, even though the p level is .04. Now lets try a 90% interval.

```

|Confidence Interval Input Values|
|Number of observations:          62
|K (Number of variables, including criterion): 2
|R Squared:                      .0676
|Confidence Level:              .9
|Go <Begin Computing>

```

```

|.9 Confidence Interval Results
|Lower Limit: 0.00145
|Upper Limit: 0.19643
|Lower Bound: 0.00930
|Plevel :    0.04127

```

That is more like it. Note that the lower limit from the 90% interval is the same as the "lower bound" from the 95% interval.

Confidence Interval for ρ , Regression Analysis. If you consider your predictor variable to be fixed rather than random (that is, you arbitrarily chose the values of that variable, or used the entire population of possible values, rather than randomly sampling values from a population of values), then the confidence interval for r^2 is computed somewhat differently. The SAS program [Conf-Interval-R2-Regr](#) can be employed to construct such a confidence interval. If you are using SPSS, see CI-R2-SPSS at my [SPSS Programs Page](#).

Bias in the Point Estimate of ρ^2 . Sample R^2 tends to overestimate population ρ^2 . When the numerator df are large, this can result in the production of a confidence interval that excludes the sample R^2 . This should not happen if you use the shrunken R^2 as your point estimate. For more details, see [Conf-Interval-R2-Regr-Warning](#).

Common Language Effect Size Statistic. Dunlap (1994, *Psychological Bulletin*, 116: 509-511) has extended the **CL** statistic to bivariate normal correlations. Assume that we have randomly sampled two individuals' scores on X and Y. If individual 1 is defined as the individual with the larger score on X, then the **CL** statistic is the probability that individual 1 also has the larger score on Y. Alternatively the **CL** statistic here can be interpreted as the probability that an individual will be above the mean on Y given that we know e is above the mean on X. Given r , $CL = \frac{\sin^{-1}(r)}{\pi} + .5$. Dunlap uses Karl Pearson's (1896) data on the correlation between fathers' and sons' heights ($r = .40$). **CL** = 63%. That is, if Joe is taller than Sam, then there is a 63% probability that Joe's son is taller than Sam's son. Put another way, if Joe is taller than average, then there is a 63% probability that Joe's son is taller than average too. Here is a little table of **CL** statistics for selected values of r , just to give you a feel for it.

r	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95	.99
CL	50%	53%	56%	60%	63%	67%	70%	75%	80%	86%	90%	96%

Multiple Correlation and Associated Statistics

The programs mentioned above can be used to put confidence intervals on multiple R^2 too. Cohen employed the effect size statistic f^2 . For the overall model, $f^2 = \frac{R^2}{1 - R^2}$. Cohen considered an f^2 of **.02** to be a small effect, **.15** a medium effect, and **.35** a large effect. We can translate these values of f^2 into proportions of variance by dividing f^2 by $(1 + f^2)$: A small effect accounts for 2% of the variance in the criterion variable, a medium effect accounts for 13%, and a large effect 26%.

For a squared partial correlation, Cohen employed the same definition, that is, $f^2 = \frac{pr^2}{1 - pr^2}$.

You can convert a squared semipartial coefficient to a partial coefficient this way: $pr_i^2 = \frac{sr_i^2}{1 - R_{y.12...(i)...p}^2}$

and then compute f^2 , or, more easily, compute f^2 this way: $f^2 = \frac{sr_i^2}{1 - R_{full}^2}$.

Consider this example. We are predicting graduate students' grade point average from their GRE scores (Verbal and Quantitative), Miller Analogies Test scores, and Average Rating obtained from faculty who interviewed the student prior to admission.

```

The REG Procedure
Model: MODEL1
Dependent Variable: GPA

Number of Observations Read      30
Number of Observations Used      30

Analysis of Variance

Source          DF          Sum of Squares          Mean Square          F Value          Pr > F
    
```

Model	4	6.68313	1.67078	11.13	<.0001
Error	25	3.75153	0.15006		
Corrected Total	29	10.43467			

Root MSE	0.38738	R-Square	0.6405
Dependent Mean	3.31333	Adj R-Sq	0.5829
Coeff Var	11.69147		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Squared Semi-partial Corr Type II	Squared Partial Corr Type II
Intercept	1	-1.73811	0.95074	-1.83	0.0795	.	.
GRE_Q	1	0.00400	0.00183	2.18	0.0385	0.06860	0.16023
GRE_V	1	0.00152	0.00105	1.45	0.1593	0.03027	0.07766
MAT	1	0.02090	0.00955	2.19	0.0382	0.06887	0.16076
AR	1	0.14423	0.11300	1.28	0.2135	0.02343	0.06118

For the full model R^2 , $f^2 = \frac{.6405}{1-.6405} = 1.78$, a whopper effect.

For the partial effect of GRE_Q, $f^2 = \frac{.16023}{1-.16023} = .191$, a medium partial effect.

Equivalently, using the semipartial, $f^2 = \frac{sr_i^2}{1-R_{full}^2} = \frac{.0686}{1-.6405} = .191$.

One Way Analysis of Variance, Independent Samples

Consider this source table:

Source	SS	df	MS	F
Teaching Method	130	3	43.33	86.66
Error	8	16	0.50	
Total	138	19		

Strength of Effect Estimates – Proportions of Variance Explained

This ANOVA is really just a multiple regression analysis with $k-1$ dummy variables, where k is the number of groups. Not surprisingly, the most commonly used strength of effect estimates are essentially R^2 statistics.

The η^2 statistic can be computed as $\frac{SS_{AmongGroups}}{SS_{Total}}$ from the ANOVA source table. This provides a fine measure of the strength of effect of the classification variable in the sample data, but it

generally overestimates the population η^2 . My program [Conf-Interval-R2-Regr.sas](#) will compute an exact confidence interval about eta-squared. If you are using SPSS, see CI-R2-SPSS at my [SPSS Programs Page](#). For our data $\eta^2 = 130/138 = .94$. A 95% confidence interval for the population parameter extends from .84 to .96.

One well-known alternative is **omega-squared**, ω^2 , which estimates the proportion of the variance in Y in the population which is due to variance in X, and is less biased than is η^2 .

$$\omega^2 = \frac{SS_{Among} - (K-1)MS_{Error}}{SS_{Total} + MS_{Error}}. \text{ For our data, } \omega^2 = \frac{130 - (3) \cdot .5}{138 + .5} = .93.$$

Cohen's f

Cohen's f (effect size) is computed as $\sqrt{\frac{\sum_{j=1}^k \sum (\mu_j - \mu)^2}{k\sigma_{error}^2}}$, where μ_j is the population mean for a single group, μ is the grand mean, k is the number of groups, and error variance is the mean within group variance. This can also be computed as $\frac{\sigma_{means}}{\sigma_{error}}$, where the numerator is the standard deviation of the population means and the denominator is the within-group standard deviation. We assume equal sample sizes and homogeneity of variance.

Suppose that the effect size we wish to use is one where the three populations means are 480, 500, and 520, with the within-group standard deviation being 100. Using the first formula above,

$$f = \sqrt{\frac{400 + 0 + 400}{3(100)^2}} = .163. \text{ Using the second formula, the population standard deviation of the}$$

means (with k , not $k-1$, in the denominator) is 16.33, so $f = 16.33 \div 100 = .163$. By the way, David Howell uses the symbol ϕ' instead of f .

Eta-Squared

You should be familiar with η^2 as the treatment variance expressed as a proportion of the total variance. If η^2 is the treatment variance, then $1-\eta^2$ is the error variance. With this in mind, we can

define $f = \sqrt{\frac{\eta^2}{1-\eta^2}}$. Accordingly, if you wish to define your effect size in terms of proportion of variance explained, you can use this formula to convert η^2 into f .

Cohen considered an f of **.10** to be a small effect, **.25** a medium effect, and **.40** a large effect. Rearranging terms in the previous formula, $\eta^2 = \frac{f^2}{1+f^2}$. Using this to translate Cohen's guidelines into proportions of variance, a small effect is one which accounts for about 1% of the variance, a medium effect 6%, and a large effect 14%.

Benchmarks for η^2 .

- .01 (1%) is small but not trivial
- .06 is medium

- .14 is large

A Word of Caution. Rosenthal has found that most psychologists misinterpret strength of effect estimates such as r^2 and ω^2 . Rosenthal (1990, *American Psychologist*, 45, 775-777.) used an example where a treatment (a small daily dose of aspirin) lowered patients' death rate so much that the researchers conducting this research the research prematurely and told the participants who were in the control condition to start taking a baby aspirin every day. So, how large was the effect of the baby aspirin? As an odds ratio it was 1.83 – that is, the odds of a heart attack were 1.83 times higher in the placebo group than in the aspirin group. As a proportion of variance explained the effect size was .0011 (about one tenth of one percent).

One solution that has been proposed for dealing with r^2 -like statistics is to report their square root instead. For the aspirin study, we would report $r = .033$ (but that still sounds small to me).

Also, keep in mind that anything that artificially lowers “error” variance, such as using homogeneous subjects and highly controlled laboratory conditions, artificially inflates r^2 , ω^2 , etc. Thus, under highly controlled conditions, one can obtain a very high ω^2 even if outside the laboratory the IV accounts for almost none of the variance in the DV. In the field those variables held constant in the lab may account for almost all of the variance in the DV.

What Confidence Coefficient Should I Employ for η^2 and *RMSSE*?

If you want the confidence interval to be equivalent to the ANOVA *F* test of the effect (which employs a one-tailed, upper tailed, probability) you should employ a confidence coefficient of $(1 - 2\alpha)$. For example, for the usual .05 criterion of statistical significance, use a 90% confidence interval, not 95%. Please see my document at <http://core.ecu.edu/psyc/wuenschk/docs30/CI-Eta2-Alpha.doc>.

Strength of Effect Estimates – Standardized Differences Among Means

When dealing with differences between or among group means, I generally prefer strength of effect estimators that rely on the standardized difference between means (rather than proportions of variance explained). We have already seen such estimators when we studied two group designs (Hedges' *g*) – but how can we apply this approach when we have more than two groups?

My favorite answer to this question is that you should just report estimates of Cohen's *d* for those contrasts (differences between means or sets of means) that are of most interest – that is, which are most relevant to the research questions you wish to address. Of course, I am also of the opinion that we would often be better served by dispensing with the ANOVA in the first place and proceeding directly to making those contrasts of interest without doing the ANOVA.

There is, however, another interesting suggestion. We could estimate the average value of Cohen's *d* for the groups in our research. There are several ways we could do this. We could, for example, estimate *d* for every pair of means, take the absolute values of those estimates, and then average them.

James H. Steiger (2004: *Psychological Methods*, 9, 164-182) has proposed the use of *RMSSE* (root mean square standardized effect) in situations like this. Here is how the *RMSSE* is calculated:

$$RMSSE = \sqrt{\left(\frac{1}{k-1}\right) \sum_1^k \left(\frac{M_j - GM}{\sqrt{MSE}}\right)^2}$$

, where *k* is the number of groups, *M_j* is a group mean, *GM* is the overall (grand) mean, and the standardizer is the pooled standard deviation, the square root of the within groups mean square, *MSE* (note that we are assuming homogeneity of variances). Basically

what we are doing here is averaging the values of $(M_j - GM)/SD$, having squared them first (to avoid them summing to zero), dividing by among groups degrees of freedom $(k - 1)$ rather than k , and then taking the square root to get back to un-squared (standard deviation) units.

Since the standardizer (sqrt of MSE) is constant across groups, we can simplify the expression

above to
$$RMSSE = \sqrt{\left(\frac{1}{k-1}\right) \frac{\sum(M_j - GM)^2}{MSE}} .$$

For our original set of data, the sum of the squared deviations between group means and grand mean is $(2-5)^2 + (3-5)^2 + (7-5)^2 + (8-5)^2 = 26$. Notice that this is simply the among groups sum

of squares (130) divided by n (5). Accordingly, $RMSSE = \sqrt{\left(\frac{1}{4-1}\right) \frac{26}{.5}} = 4.16$, a Godzilla-sized average standardized difference between group means.

We can place a confidence interval about our estimate of the average standardized difference between group means. To do so we shall need the NDC program from Steiger's page at <http://www.statpower.net/Content/NDC/NDC.exe>. Download and run that exe. Ask for a 95% CI and give the values of F and df .

Click "COMPUTE."

You are given the CI for lambda, the noncentrality parameter:

Now we transform this confidence interval to a confidence interval for RMSSE by with the following transformation (applied to each end of the *CI*): $RMSSE = \sqrt{\frac{\lambda}{(k-1)n}}$. For the lower boundary, this yields $\sqrt{\frac{102.646301}{(3)5}} = 2.616$, and for the upper boundary $\sqrt{\frac{480.288073}{(3)5}} = 5.659$.

That is, our estimate of the effect size is between King Kong-sized and Godzilla-sized.

Steiger noted that a test of the null hypothesis that Ψ (the parameter estimated by RMSSE) = 0 is equivalent to the standard ANOVA *F* test if the confidence interval is constructed with 100(1-2 α)% confidence. For example, if the ANOVA were conducted with .05 as the criterion of statistical significance, then an equivalent confidence interval for Ψ should be at 90% confidence -- Ψ cannot be negative, after all. If the 90% confidence interval for Ψ includes 0, then the ANOVA *F* falls short of significance, if it excludes 0, then the ANOVA *F* is significant.

Factorial Analysis of Variance, Independent Samples

Eta-squared and **omega-squared** can be computed for each effect in the model. For η^2 , simply divide the effect sum of squares by the total (corrected) sum of squares. With omega-squared, substitute the effect's *df* for the term $(k-1)$ in the formula shown earlier for the one-way design.

Partial Eta-Squared. The value of η^2 for any one effect can be influenced by the number of and magnitude of other effects in the model. Suppose we were conducting research on the effect of sex (female, male) and type of experimental therapy on a health variable. We wish to generalize our results to the population of persons for whom the experimental therapy might be useful. Type of therapy is not a variable in the general population, since it is experimental and is only being used in this study. Accordingly, when estimating the magnitude of effect for the sex variable, we might elect

to exclude from the denominator of eta-squared the variance due to type of therapy and its interaction with sex. The resulting statistic is called partial eta-squared, $\eta_p^2 = \frac{SS_{Effect}}{SS_{Effect} + SS_{Error}}$. Of course, this will make partial eta-squared larger than regular eta-squared. The question answered by partial eta-squared is this: Of the variance that is not explained by other effects in the model, what proportion is explained by this effect.

For Sex x Therapy analysis, when estimating the magnitude of effect for type of therapy and its interaction with sex, it would not be reasonable to exclude from the denominator the variance due to sex, since sex is variable in the population of interest.

Partial eta-square values can be considerably larger than the eta-square or omega-square values. Clearly this statistic can be used to make a small effect look moderate in size or a moderate-sized effect look big. It is even possible to get partial eta-square values that sum to greater than 100%. That makes me a little uncomfortable. Even more disconcerting, many researchers have incorrectly reported partial eta-squared as being regular eta-squared. Pierce, Block, and Aguinis (2004, *Educational and Psychological Measurement*, 64, 916-924) found articles in prestigious psychological journals in which this error was made. Apparently the authors of these articles (which appeared in *Psychological Science* and other premier journals) were not disturbed by the fact that the values they reported indicated that they had accounted for more than 100% of the variance in the outcome variable – in one case, the authors claimed to have explained 204%. Oh my.

Eta-Squared or Partial Eta-Squared? Which one should you use? I am more comfortable with eta-squared, but can imagine some situations (like the Sex x Therapy example above) where the use of partial eta-squared might be justified. Kline (2004, *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association. 325 pp.) has argued that when a factor like sex is naturally variable in both the population of interest and the sample, then variance associated with it should be included in the denominator of the strength of effect estimator, but when a factor is experimental and not present in the population of interest, then the variance associated with it may reasonably be excluded from the denominator.

To learn more about this controversial topic, read Chapter 7 in Kline (2004). You can find my notes taken from this book on my Stat Help page, [Beyond Significance Testing](#).

Constructing a Confidence Interval for Eta-Squared or Partial Eta-Squared. As was the case with one-way ANOVA, one can use my program [Conf-Interval-R2-Regr.sas](#) to put a confidence interval about eta-square or partial eta-square. You will, however, need to compute an adjusted F when putting the confidence interval on η^2 . See my document [Two-Way Independent Samples ANOVA on SAS](#).

Contingency Table Analysis

I find **phi** an appealing estimate of the magnitude of effect of the relationship between two dichotomous variables and **Cramér's phi** appealing for use with tables where at least one of the variables has more than two levels.

The phi coefficient is, quite simply, the Pearson r computed between two dichotomous variables. When using SAS or SPSS you can request the phi and Cramer Cramér's phi statistics. Cramér's phi does not suffer from the problems common to other strength of effects statistics for

contingency tables – for example, the value of Cramér’s phi is not influenced by the dimensionality of the table (2 x 2, 2 x 3, 3 x 3, etc.), but other strength of effect estimates are.

For 2 x 2 tables, I generally prefer **odds ratios**. Consider the results of some of my research on attitudes about animals (Wuensch, K. L., & Poteat, G. M. Evaluating the morality of animal research: Effects of ethical ideology, gender, and purpose. *Journal of Social Behavior and Personality*, 1998, 13, 139-150. Participants were pretending to be members of a university research ethics committee charged with deciding whether or not to stop a particular piece of animal research which was alleged, by an animal rights group, to be evil. After hearing the evidence and arguments of both sides, 140 female participants decided to stop the research and 60 decided to let it continue. That is, the odds that a female participant would stop the research were $140/60 = 2.33$. Among male participants, 47 decided to stop the research and 68 decided to let it continue, for odds of $47/68 = 0.69$. The ratio of these two odds is $2.33 / .69 = 3.38$. In other words, the women were more than 3 times as likely as the men to decide to stop the research.

With the appropriate software, it is easy to construct a confidence interval for an odds ratio. For example, with SPSS simply do the analysis as a binary logistic regression and ask for confidence intervals for the odds ratios.

Benchmarks.

Phi	Odds Ratio*	Size
0.1	1.5	Small, not trivial
0.3	3.5	Medium
0.5	9	Large

*For a 2 x 2 table with both marginals distributed uniformly.

For examples of contingency tables with small, medium, and large effects, see my document [Estimating the Sample Size Necessary to Have Enough Power](#). Please note that the correspondence between phi and odds ratios in the table above assumes that the marginals are uniform. For a fixed value of the odds ratio, the more the marginals deviate from uniform the lower the phi. See [this document](#) for more details.

Multivariate Statistics

Most multivariate analyses provide statistics very similar to r^2 and η^2 . For example, in a canonical correlation/regression analysis one obtains squared canonical correlations for each root. For each root the canonical correlation is simply the correlation between a weighted linear combination of the Y variables and a weighted linear combination of the X variables.

MANOVA and DFA. For each root (canonical variate, discriminant function) one can easily obtain the squared canonical correlation. This is absolutely equivalent to the η^2 that would be obtained were you to compute the canonical variate or discriminant function scores for each subject and then conduct an ANOVA comparing the groups on the canonical variate or discriminant function. From that ANOVA the eta-squared would be computed as $\frac{SS_{\text{among_groups}}}{SS_{\text{total}}}$.

In MANOVA and DFA we may test our treatment by finding the ratio of the determinant (generalized variance) of the error SSCP matrix to the determinant of the sum of the treatment and error SSCP matrices. This ratio is called Wilks' Lambda (Λ). Since the ratio is $\frac{\text{error}}{\text{error} + \text{treatment}}$,

one may interpret the Wilks' Lambda as the proportion of variance in the DVs that is not accounted for by the IV. An **eta-squared** statistic can be computed as $\eta^2 = 1 - \Lambda$. We may interpret this statistic as being the proportion of variance in the continuous variables that is accounted for by the classification variable. When there are only two treatment groups (and thus only one root), this η^2 will equal the squared canonical correlation.

Binary Logistic Regression. For the overall model, the **Cox & Snell R^2** can be interpreted like R^2 in a multiple regression, but cannot reach a maximum value of 1. The **Nagelkerke R^2** can reach a maximum of 1. Classification result statistics (total percentage correct classifications, sensitivity, specificity, false positive and false negative error rates) also speak to magnitude of effect. The strength of the partial effect of individual predictors is measured by odds ratios.

It may be useful to standardize continuous predictor variables prior to computing odds ratios. Consider recent research showing the relationship between retention in ECU's engineering program and three continuous predictors: high school GPA, quantitative SAT, and score on the Openness scale of a Big Five personality test.

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a HSGPA	1.296	.506	6.569	1	.010	3.656
SATQ	.006	.003	4.791	1	.029	1.006
Openness	.100	.045	4.832	1	.028	1.105
Constant	-10.286	2.743	14.063	1	.000	.000

a. Variable(s) entered on step 1: HSGPA, SATQ, Openness.

Look at the odds ratios [Exp(B)]. One might be fooled into thinking that the effect of HSGPA is very much greater than that of SATQ and Openness, but one need consider the scale of measurement of the predictors. For each one point increase in HSGPA the odds of being retained are multiplied by 3.656. For each one point increase in SATQ the odds are multiplied by 1.006 – but a one point increase in GPA is a very large increase in GPA, and a one point increase in SATQ is a very small increase.

Now look at the odds ratios after standardizing the predictors.

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a ZHSGPA	.510	.199	6.569	1	.010	1.665
ZSATQ	.440	.201	4.791	1	.029	1.553
ZOpenness	.435	.198	4.832	1	.028	1.545
Constant	-.121	.188	.414	1	.520	.886

a. Variable(s) entered on step 1: ZHSGPA, ZSATQ, ZOpenness.

Aha, now it is clear that the three predictors have similar unique contributions to the model.

Why Confidence Intervals?

It is not common practice to provide confidence intervals about point estimates of effect size estimates, but it should be. Why is it not adequate to provide just a test of a null hypothesis and an associated point estimate of effect size? My answer to this question is that the p value and the point estimate tell you nothing about the precision with which you have estimated the strength of effect. Consider the following possible outcomes, where the confidence intervals are given for r or an r -like strength of effect estimator:

Significant Results

- $CI = .01, .03$ – Although we can be confident of the direction of the effect, we can also be confident that the size of the effect is so small that it might as well be zero. “Significant” in this case is a very poor descriptor of the effect.
- $CI = .02, .84$ – We can be confident of the direction of the effect and that the effect is probably large enough to be of importance, but we have estimated it with very little precision – it could be trivially small or humongous. We really need to get more data so we can obtain a more precise estimate of the size of the effect.
- $CI = .51, .55$ – We can be quite confident that we have an effect that would, in the behavioral sciences, be considered large (in most contexts).

Not Significant Results

- $CI = -.46, +.43$ – About all this tells us is that the researcher needs a lot more data than he has on hand.
- $CI = -.74, +.02$ – Although we cannot be confident about the direction of the effect, I’d bet it is negative. Clearly we need more data.
- $CI = -.02, +.01$ – This is actually a very impressive confidence interval, because it tells us that the size of the effect is so small as to be trivial. Although “not significant,” this CI is telling us pretty much the same thing that the significant CI of $.01, .03$ told us. Suppose that the two variables were type of drug (brand name or generic) and physiological response. Knowing with confidence that the association is essentially zero is very useful information.

[Return to Wuensch’s Statistics Help Page](#)

[Fair Use of This Document](#)