Sleep and Hypnosis
A Journal of Clinical Neuroscience and Psychopathology

Sleep Hypn.2019 Dec;21(4): 339-350
http://dx.doi.org/10.37133/Sleep.Hypn.2019.21.0207

ORIGINAL ARTICLE

Reliability, Validity, Effect Sizes and Confidence Intervals in Hypnosis Research

Marty Sapp[1]

1 University of Wisconsin-Milwaukee

ABSTRACT
Hypnosis researchers need to provide effect size measures, and that they need to calculate reliability indices for their data. In addition, hypnosis researchers need to think meta-analytically, and not to apply mindlessly statistics and measurement (Fidler, Cumming, Thomason, Pannuzzo, Smith, Fyffe, et al 2005). Moreover, confidence intervals are needed within hypnosis research. Finally, this article described applications of reliability, validity, effect sizes, and confidence intervals to hypnosis research.
Keywords: reliability, validity, effect sizes, confidence intervals, hypnosis research.

History of effect sizes

Material in this paper is based on Sapp (2017), and he provides a primer on effect sizes, simple research designs, and confidence intervals. Huberty (2002) found that the history of effect size started around 1940. The correlation ratio or eta coefficient was proposed during the 1940s. The correlation ratio is used to measure curvilinear relationships. In addition, eta measures the relationship between a grouping variable and a dependent or outcome variable. During this period, eta squared was connected to analysis of variance (ANOVA) to show the variance accounted for on a dependent variable. Eta=.826. Cohen characterized eta squared of .01 as a small effect size, an eta squared of .06 as a medium effect size, and an eta squared of .14 as a large effect size.

The .683 is the variance accounted for on the dependent variable, and .826 is the correlation of the group identifications with the dependent variable. Ronald A. Fisher (1890-1961), in 1924, derived the probability of eta in the context of ANOVA. Truman (1935) Kelley (1884-1961) proposed an adjustment to the eta squared within the context of ANOVA. Some statisticians refer to this as the partial eta squared. The psychologist William L. Hays (1926-1995) in his popular textbook proposed omega squared as an alternative to eta squared. Omega squared is said to be derived through unbiased estimates. Omega squared =SSB-(K-1)MSW/(SST+MSW). Where SSB equals the sum of squares between and K equals the number of groups. MSW is the mean squares within, and SST is the total sum of squares. Generally, omega squared and eta squared will not differ much. If the levels of the grouping variable (independent variable) are random, in contrast to being fixed, the intraclass

***Correspondence:** sapp@uwm.edu, +(414)229-4599, University of Wisconsin-Milwaukee, Department of Educational Psychology, USA.
Received: 30 November 2018 **Accepted:** 29 October 2019

Sleep and Hypnosis
Journal homepage:
http://www.sleepandhypnosis.org
ISSN:1302-1192 (Print) 2458-9101 (Online)

correlation coefficient can be used as an effect size. The formula for the intraclass correlation R is the following:

R=(MSB-MSW)/[MSB+(n-1)MSW]

MSB and MSW are the numerator and denominator from an F statistic or test and n equals the number of participants per group.

In summary, at least three strengths of relationship effect sizes were proposed between 1935 to 1963: eta squared, omega squared, and the intraclass correlation coefficient. Karl Pearson (1857-1936), in 1910, proposed the biserial correlation coefficient. It is used when a continuous variable is forced into a discrete variable and is correlated with a continuous variable. For example, suppose we were interested in the correlation between hypnotizablity and creative imagination. Both of these variables are continuous, but we forced the hypnotizability scores into high and low hypnotizability. The correlation between these two variables would be the biserial correlation coefficient. The biserial correlation coefficient cannot be used in regression in order to predict y values or dependent variables. Also, confidence intervals cannot be placed around the biserial correlation coefficient. Finally, the biserial correlation coefficient is less reliable than the Pearson correlation coefficient, and it is not recommended as an effect size.

Jacob Cohen, in 1969, proposed an effect size for a two group mean comparison, and Huberty (2002) referred to these as group differences indices. Cohen defined his effect size as the differences between means divided by the pool standard deviation across the two groups. Like Cohen, the statistician Gene V. Glass also proposed a d effect size as the differences between means divided by the control group standard deviation. In addition, the statistician Larry V. Hedges took exception with Cohen and Glass, and he proposed an adjusted d that he called g (Huberty, 2002). Cohen also proposed a standard difference type of effect size for multiple groups or multiple means context (ANOVA), and he used the letter f as this effect size, and it is the following formula:

$f=[(K-1)F/N]^{1/2}$

K is the number of groups, and F is the F statistics from ANOVA. N is the total group size. When using Cohen's power tables the average group size is used or the harmonic mean if the group sizes are unequal. F can be seen as the standard deviation of the standardized means, or the variability of the group means relative to the standard deviation (Huberty, 2002). Cohen (1977; 1988) characterized f equals .10 as a small effect size, f=.25 as a medium effect size, and f>.40 as a large effect size.

Huberty (2002) discussed another effect size based on overlap indices. Within a two-group situation, if two have a large amount of overlap the effect size will be small. Cohen (1988) also defined d as the percent of non-overlap of the treatment group scores with those of the untreated group. An effect size of zero indicates that the distribution of scores of the treatment group overlap completely with the distribution of the control group. Cohen (1977) provided the following rough guidelines for interpreting the d effect size: **d = .2 small effect size, d = .5 medium effect size, and d = .8 large effect size.**

One should not just blindly accept the standards based on Cohen's work, but interpret effect sizes within a given professional area.

The r effect size and d effect size are related in that

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Also, d can be expressed as t using the following formula: $d=t(1/n1+1/n2)^{1/2}$. The t is the value from a t test, and the n1 and n2 are the respective group sizes.

The following is the relationship between r and d:

d effect size	r effect size
4.0	.894
3.8	.885
3.6	.874
3.4	.862
3.2	.848
3.0	.832
2.8	.814
2.6	.793
2.4	.768
2.2	.740
2.0	.707
1.9	.689
1.8	.669
1.7	.648
1.6	.625
1.5	.600
1.4	.573
1.3	.545
1.2	.514
1.1	.482
1.0	.447
.9	.410

large	**.8**	**.371**
	.7	.330
	.6	.287
medium	**.5**	**.243**
	.4	.196
	.3	.148
small	**.2**	**100**
	.1	.050
	0	0

With his history of effect sizes, the final group of effect sizes that he discussed were the multivariate indices (Huberty, 2002). The concept of multiple regression or the multiple correlation coefficient was developed in 1914 by Pearson and Lee (1897). Cohen's f^2 equals $R^2/(1-R^2)$. R is the multiple correlation coefficient. Multivariate Analysis of Variance (MANOVA) is applicable to a group variable situation where participants are measured on two or more dependent or outcome variables. Maurice M. Tatsuoka (1922-1996) summarized the literature in this area in 1973. Tatsuoka (1970) connected Samuel S. Wilks' (1906-1964) Lambda to the MANOVA context as a measure of multivariate strength of association. The smaller the value of Wilks' Lambda, the stronger the multivariate effect.

After reviewing several journals within the area of hypnosis, I found few studies addressing basic measurement issues, effect sizes, and confidence intervals. With this is in mind, the purpose of this section is to address these factors. Because sufficient narrative is used in place of formulas, I hope that researchers can apply these concepts to their research. These sections are divided into the following parts: reliability, effect sizes, definition of multivariate statistics effect size, testing calculated validity coefficients against hypothesized values, standard error of estimate, confidence intervals around validity, and discussion.

Reliability

Classical test theory is the model often taught in basic psychological measurement classes. Measurement experts have used this theory of measurement since the turn of the century. Many times, it is used to find reliability measures such as test-retest, internal consistency, and so on. It is also referred to as the true score, and it has the following mathematical model:

X = T + E

X = a person's score or an observed score

T = a person's true score

E = the error score

Theoretically, reliability can be expressed as the ratio of true score variance divided by the observed score variance. If we symbolize reliability as r_{xx}, it can be expressed mathematically as

$$r_{xx} = \frac{S_t^2}{S_x^2} = \frac{\text{true score variance}}{\text{observed score variance}}$$

A specific form of reliability, called coefficient alpha, is defined by two quantities. First, the number of test items divided by the number of test items minus one. The second quantity is one minus the sum of item variances divided by the total test variance. Finally, these quantities are multiplied. In summary, coefficient alpha, like other forms of reliability fits the definition involving variances. The following is the formula for coefficient alpha:

$$K / (K-1)\left[1 - \frac{\sum S_i^2}{S_t^2}\right]$$

where K equals the number of items

ΣS^2_i equals the variance across-test items

S^2_t equals the variance for the participants' total test scores.

Verbally, coefficient alpha is the following:

$$\frac{\text{No. of Items}}{\text{No. of Items - 1}}\left[1 - \frac{\text{Sum of Item Variances}}{\text{Test Variance}}\right]$$

Sapp (2013) recommended interpreting test scores using the standard error of measurement. This index measures the amount of error within test items. Essentially, this is the standard deviation for a set of items, and this formula is the following:

$$S_e = S_x\sqrt{1 - r_{xx}}$$

Reliability is the variance that is accounted for on a set of hypnotizability test items; hence, it is a squared correlation or squared area. Reliability is the percent of variance accounted for on a hypnosis measure. Once a hypnosis test is standardized, the reliability that is reported in a manual is the reliability measure for the standardization sample, but this value does not tell one how another independent sample will respond to those test items; therefore, within the 21st century, measurement theorists make a distinction between reliability of the standardization sample and reliability of an independent sample. The important point is that reliability involves how individuals respond to hypnosis test items; hence, reliability is not invariant; meaning it does not change from sample to sample. The only way to know reliability for a given sample is to calculate it. In essence, reliability is the consistency that a sample responds to a set of test items. Sadly, within a multicultural perspective, often minorities are not included within the standardization process for available hypnosis tests.

Internal consistency can determine the consistency of test items. Coefficient alpha is the most used measure of internal consistency. Suppose that 12 Latino college students completed 6 items that measured dissociation, and these items were rated on a 4 point scale. The following output from SPSS has coefficient alpha, and the 95% confidence interval around the population coefficient alpha. These data for this example are the following:

These data that follow are in the following form: participant, item 1, item 2, item 3 ,item 4, item5, and item 6:

1.	1	2	3	1	4	1
2.	1	1	1	1	1	1
3.	1	1	2	2	4	2
4.	3	3	3	3	3	3
5.	1	2	3	4	4	2
6.	1	2	3	4	4	1
7.	2	1	3	3	3	4
8.	2	1	4	4	4	1
9.	2	1	3	4	3	2
10.	2	2	4	3	3	3
11.	2	2	3	3	3	3
12.	2	2	4	3	3	4

The SPSS control lines for these data are the following:

Reliability

/VARIABLES=item1 item2 item3 item4 item5 item6
/SCALE('ALL VARIABLES') ALL/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE CORR ANOVA
/ICC=MODEL(MIXED) TYPE(CONSISTENCY)
CIN=95 TESTVAL=0 .

The following are the output for coefficient alpha:

Reliability statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.691	.709	6

Coefficient alpha or Cronbach's alpha was .691. This tell us that 69.1% on these items is true score variance, and 1-.691, and .309 or 30.9 % is the error variance. Is summary, the point estimate, or alpha for this sample data was .691. Later, when confidence intervals are discussed, a confidence interval will be provided for the population coefficient alpha.

Validity

Validity, determines if hypnosis items measure what they are suppose to measure. Like reliability, since minorities are seldom included within standardization samples for hypnosis tests that measure validity. Often, when researchers speak about validity, they are referring to criterion validity. Criterion validity tells the degree that items from two hypnosis tests correlate. Sapp (2006) reported that validity coefficients tend to fall with .20 and .60. Unlike reliability coefficients, validity coefficients must be squared to find the variance account for, or the coefficient of determination. For example, a validity coefficient of .5 states that .25 or 25% of the variance can be explained, and 75% of the variance is unexplained.

Effect sizes

Effect sizes are seldom reported within hypnosis research. Effect sizes allow researchers to if statistical results have practical significance, and they allow one to determine the degree of effect hypnosis has within

a population; or simply stated, the degree in which the null hypothesis may be false. There are over 40 different effect sizes, but, as I discussed within the history of effect size section, they can be group into two broad areas-means differences effect sizes like the d effect size and correlational effect sizes like effect size r (Ferguson, 2009).

Cohen (1977) defined the most basic effect measure, the statistic that is synthesized, as an analog to the t-tests for means. Specifically, for a two-group situation, he defined the d effect size as the differences between two population means divided by the standard deviation of either population, because homogeneity or equality of variances is assumed. This effect size has the general formula:

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

μ_1 = the treatment group

μ_2 = the control group

σ = population standard deviation

Suppose μ_1 equals the population mean, and in this case we are using it to represent the treatment group population mean of 1.0. And let us assume that μ_2, the population mean for the control group, equals .2, and, finally, σ = 1.00. By substitution,

$$d = \frac{1.0 - .2}{1} = .8$$

Hedges' g or d is Cohen's d[1-(3/4df-1)] The df is the total sample size minus two. Now, this formula can create confusion because different effect sizes can be found for the same data. For example, the σ or population standard deviation can be from the control group posttest measure. In addition, it can be the pretest standard deviation for the control group, or it can be the pooled or weighted standard deviation that involves both groups. Therefore, within a study, at least three different d effect sizes can be obtained. First, one based on the control group posttest measure standard deviation. Second, another based on the control group pretest measure, and a d effect size measure based on the average standard deviation for the treatment and control group.

As stated early, the differences between means divided by the control group standard deviation is actually Glass's delta and not the specific d that Cohen proposed, but many researchers assume that all ds are Cohen's ds. The d that Cohen proposed was the differences between two group means divided by the pool standard deviations of the treatment group and control group. Cohen's d assumes homogeneity of variance and this assumption is violated one would want to chose which standard deviation to use because they cannot be pooled. With repeated measures designs, it can be argued that it does not make sense to calculate Cohen's d. This is because Cohen's d was developed for independent groups. One can use eta squared with repeated measures designs. In practice, many researchers, for Cohen's d, find the differences between the treatment and control groups means divided by the average of the standard deviation of the treatment group and the standard deviation of the control group. Cohen's d is upward biased and this is why Hedges developed his d to take into account this biased. Cohen's d is more appropriate for population data while Hedges' d is more appropriate for sample data. These d effect sizes from several studies can be averaged, and the result is an overall effect size for a series of studies. Meta-analysis is just the overall effect for a given area or mean effect size, and it is obtained by adding the effects sizes and dividing by the total number (Ferguson, 2007). Also, effect sizes are used for statistical power analysis, or the probability of rejecting a false null hypothesis.

Although Cohen (1977) provided the following rough guidelines for interpreting the *d* effect size: *d* = .2 small effect size, *d* = .5 medium effect size, and *d* = .8 large effect size, researchers should not interpret blindly effect sizes as small, medium, and large. One must interpret effect sizes within a given professional area.

There is another effect size called *r*, and it was described by Rosenthal (1984). The reader may remember that *r* is the Pearson product-moment correlation coefficient. Mathematically, *r* is the covariance, the amount two variables vary co-varies,

divided by the number of pairs times the product of the standard deviation for each variable. The following is the formula for the

Pearson product-moment correlation: $r = \frac{\Sigma Z_x Z_y}{N}$.

Here, Z_x is every X value minus the mean of the X values divided by the standard deviation of the X values. Similarly, Z_y is every Y value minus the mean of the Y values divided by the standard deviation of the Y values. Z_x and Z_y are analogous to standard deviation and are referred to as moments, hence the name Pearson Product-Moment Correlation. The reader should note that a moment is a measure of variability like the standard deviation. Like Cohen (1977), Rosenthal (1984) provided the following rough guidelines for *r*: *r* = .1 small effect size, *r* = .3 medium effect size, and *r* = .5 large effect size. The following section will describe a common multivariate effect size that is analogous to the d effect size.

Definitions of multivariate statistics

The term multivariate can be a confusing term, but in one sense it involves examining several variables simultaneously. Within a regression context, it is the relationship between two or more predictors (independent variables) and a dependent variable. From a multivariate regression context, it involves the relationship between two or more predictors and two or more dependent variables. Other multivariate correlation methods are path analysis, factor analysis, principal components analysis, canonical correlation, and predictive discriminant analysis (Stevens, 2002).

When two or more groups of participants are measured on several dependent variables, this is a multivariate analysis of variance (MANOVA), a multivariate extension of ANOVA. Multivariate analysis of covariance (MANCOVA), a multivariate generalization of ANCOVA, step down analysis, a multivariate test procedure that focuses on the ordering of dependent variables through a series of analyses of covariance, and descriptive discriminant analysis, a multivariate technique that determines group membership, and log linear analysis, an extension of the chi-square test to three or more variables are all examples of multivariate statistics. In summary, if participants are measured on two or more dependent variables, a multivariate situation exists.

Why are multivariate statistics important? First, they control type I error, but with many univariate tests it cannot be easily estimated. Second, univariate statistics do not take into account the correlations among variables. Finally, multivariate statistics are more powerful statistically than univariate statistics.

Hotelling's T squared is the squared multivariate generalization of the t test. The univariate t test is the following:

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Hoteling's T^2 is the following:

$$T^2 = \frac{n_1 n_2}{n_1+n_2}\left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)$$

$\left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)'$ transpose of vector of means

S- sample covariance matrix

$\mathbf{S}^{-1}$ matrix analogue of division is inversion

$\left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)$ vectors of means

The connection between Hotelling's T^2 and F is the following:

$$F = \frac{n_1+n_2-p-1}{(n_1+n_2-2)p}T^2$$

This formula shows that T^2 provides a F distribution with p and (N-P-1) degrees of freedom. The p is the number of dependent variables and N equals the sample size. Essentially, T^2 is the comparison of between variability divided by within variability.

The univariate d and Mahalanobis distance (D^2) are the following:

univariate multivariate

$$d = \frac{\bar{y}_1 - \bar{y}_2}{s} \qquad D^2 = \left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2\right)$$

D^2 is also the following two formulas:

This is formula one. $\left[\left(n_1 + n_2\right)/ n_1 n_2\right] T^2$

This is formula two. $\frac{1}{1-r^2}\left[\frac{\left(x_{i1} - \bar{x}_1\right)^2}{s_1^2} + \frac{\left(x_{i2} - \bar{x}_2\right)^2}{s_2^2}\right.$

Formula two, clearly show how D^2 takes into account the correlation of these variables.

And T^2 is the following:

$$\left[n_1 n_2 /\left(n_1 + n_2\right)\right] D^2$$

Stevens (2002) stated that values of .25 are small effect sizes, values of .5 are medium effect sizes, and values greater than one are large effect sizes. Unlike univariate statistics, Mahalanobis distance takes into account the intercorrelation of variables. Readers can refer to Sapp, Obiakor, Gregas, and Scholze (2007) and Stevens (2002) on how to calculate Mahalanobis statistic with SPSS.

Confidence intervals

Ferguson (2009) and Sapp (2004) defined a confidence interval as an interval among an infinite number of intervals for a parameter such as population mean, population reliability coefficient, population proportion, population correlation coefficient, population difference and so on, in which one minus the alpha level would capture the population parameter a certain percentage of the time. For example, for a population mean, 95 percent of these intervals would capture the population mean and 5 percent would not. In contrast to point estimates, which describe sample data, confidence intervals describe population characteristics. More specifically, confidence intervals allow researchers to put a lower limit and upper limit around a population parameter. The 95 percent and 99 percent are most used intervals, but any interval width can be established. For the 99 percent interval, a researcher is assuming that 99% of these intervals capture these population parameters, and 1 percent would not. Clearly, a 99 percent interval is wider than a 95 percent one (Sapp, Obiakor, Scholze, & Gregas, 2007; Sapp, 2004a; Thompson, 2002).

Confidence intervals can be placed around IQ and other standardized scores. For example, the Wechsler Adults Intelligence Scale (WAIS), a commonly used measure of intelligence, has a standard error of measurement of 5. Since the standard error of measurement is interpreted in terms of the normal curve, confidence intervals can be formed around IQ scores. For example, if an African American student had an IQ score of 100 on the WAIS, this IQ scores of 100 plus and minus one times the standard error approximates the 68% confidence interval. The IQ score of 100 minus the standard error of five equals 95, which is the lower limit. And the IQ score of 100 plus 5 equals the upper limit. This means we can expect this African American student's true IQ score to fall between 95 and 105 68% of the time. Similarly, 100 plus and minus 1.96 times the standard error of measurement (5) represents the 95% confidence interval. Finally, 100 plus and minus 2.58 times the standard error of measurement forms the 99% confidence interval.

Testing calculated validity coefficients against hypothesiozed values

Just as values of reliability can be tested against hypothesized values, the same test can be performed with validity coefficients. Two independent validity coefficients can be tested for statistical significance using the following Z-test (Sapp, 1997).

$$Z = \frac{Zr_1 - Zr_2}{\left(\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}\right)^{1/2}}$$

Zr_1 and Zr_2 are Fisher's z transformations of r for the validity coefficients. Suppose that sample one had a validity coefficient or index $r_1 = .50$, with 100 participants and sample had a validity coefficient or $r_2 = .35$, also with 100 participants. The first step is to find the Fisher's z transformation for each validity coefficient. For $r_1 = .50$, the Fisher's z is $Zr_1 = .549$, and for $r_2 = .35$, the Fisher's z is .365. Substituting into the formula:

$$Z = \frac{.549 - .365}{\left(\frac{1}{97} + \frac{1}{97}\right)^{\frac{1}{2}}} = \frac{.184}{(.010309278 + .010309278)^{1/2}}$$

$$= \frac{.184}{.143591631} = 1.281411732 \text{ or } 1.28 \text{ at 2 decimal places.}$$

Because Z of 1.28 is not greater than a Z of 1.96 (critical value), the validity coefficients are not statistically significantly different. Finally, for two related or correlated validity coefficients, the formula is the following:

$$Z = \frac{Zr_1 - Zr_2}{\left(\frac{1}{N-3}\right)^{1/2}}$$

Using the validity coefficients, suppose that a population validity coefficient of .50 exists within some bivariate normal distribution, and suppose a random sample of 100 participants were drawn randomly from the normal population distribution and the sample validity coefficient is .35. Substituting into the formula provides the following:

$$Z = \frac{.549 - .365}{\left(\frac{1}{97}\right)^{1/2}} = \frac{.184}{(.010309278)^{1/2}}$$

$$= \frac{.184}{.101534617} = 1.812189836$$

= or 1.81 rounded to two decimal places.

Again, since the calculated value of Z of 1.81 is not greater than the critical value of Z, which is 1.96, the two related validity coefficients are not statistically significantly different.

Confidence intervals around validity

Confidence intervals can be placed around validity. As previously stated, validity is the correlation among a set of items that been shown to be valid with a set of items being tested to determine their validity; therefore, validity can be defined as a simple correlation. The sampling distribution of the Pearson product-moment correlation, the most commonly used one, is skewed; therefore, this correlation must be turned into a logarithmic transformation. The reader can see Sapp (2006) for these transformations.

Suppose, a researcher had a validity coefficient of .30 for a hypnosis study, how could one construct a 95% confidence interval around the population validity coefficient? First, turn the validity coefficient into its logarithmic transformation that is .31. Suppose this validity coefficient is based on 25 cases. Like the reliability example, we need the standard error, which is one divided by the square root of the number of cases minus three; therefore, the standard error is .21. The 95% confidence interval is .31 plus and minus 1.96 times .21, so the lower limit is -.10 and the upper limit is .72. We have to transform these logarithmic values back to regular correlations, and these become -.10 for the lower limit and .62 for the upper limit. The reader should notice the confidence interval -.10, .62 contains zero; therefore, the population correlation coefficient does not differ significantly from zero; therefore, there is not statistical significance.

With centralized distributions such as the normal curve and t-distribution for centralized cases, confidence intervals are straightforward. For example, the confidence interval for the one-sample t-test is the sample mean plus and minus the critical value of the t test statistics times the standard error. For the two-sample t-test case, the sample mean is replaced with the difference between means. For example, the formula for the confidence interval for a two-sample t-test is the following: $(\overline{X}_1 - \overline{X}_2) \pm (t)$ (standard error). Again, X bars are the sample means and t is the t test statistic. The t formula is the

following: $$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

The Ss squared are the standard deviations of each group squared and the Ns are the sample sizes for each group. Confidence intervals can be placed around validity indices or correlation indices and multiple squared correlations (Sapp, 2012; Steiger & Fouladi, 1997; Smithson, 2003). As previously stated, before a confidence interval can be establish, one must determine if one is working with a centralized or noncentalized distribution. The reason the normal curve is centralized is because it has a population mean of zero and a standard deviation of one. The centralized t-distribution is a generalization of the normal distribution, and it is defined by a mean of zero and degrees of freedom. Noncentalized distributions are defined by their degrees of freedom and noncentalized parameters.

The upper and lower limits for a confidence interval for a one-sample case are found by finding the mean plus and minus the critical value of the t test statistics times the standard error. The minus part of this definition provides the lower limit and the plus part provides the upper limit. The following is an example of a one-sample case with a 95 percent confidence interval.

A practical example of a one sample case 95% confidence interval

Assume that a university tested a random sample of ten students on the SAT, and the population mean was 708. The following are these students' SAT scores:

708
707
710
708
711
707
708
710
707
709

Calculate the appropriate test statistic for this design. Is the test statistic significant? Calculate a 95% confidence interval.

One-Sample Statistics				
	N	Mean	Std. Deviation	Std. Error Mean
SAT Score	10	708.5000	1.43372	.45338

One-Sample Test						
	Test Value = 708					
	T	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
SAT Score	1.103	9	.299	.50000	-.5256	1.5256

The appropriate test statistic for this design is the one-sample t-test, and statistical significance was not obtained because the level of significance or probability value was .299 for the t-test statistic; a value of .05 or lower is needed for statistical significance. The reader should notice that the 95% confidence interval of the difference between the sample mean and population mean has a lower limit of -.5256 and and upper limit of 1.5256. Since zero is included within the interval, a statistical significance difference was not found between the sample mean and the population mean. These upper and lower limits are found by taking the mean difference (sample mean of 708.5-705 =.50) plus and minus the critical value of t which is 2.262 times the standard error which is .45338. The mean difference of .50 plus 1.02554556 equals 1.5256 rounded to four decimal places. In contrast, the mean difference of .50 minus 1.02554556 equals -.5256 rounded to four decimal places. This suggests that the sample mean is representative of the population mean. The confidence interval for the one-sample case t-test equals the mean plus and minus the critical value of t times the standard error of the mean. If the population mean is not known, the sample mean alone is used to find the confidence interval. Again, the critical value of t for nine degrees of freedom is 2.262; therefore, the upper limit for this confidence interval is 708.5+2.262(.45338). 2.262(.45338)=1.02554556.

708.5+1.02554556=709.5255456 upper limit
708.5-1.02554556=707.4744544 lower limit

Finally, we are 95% confidence that the mean SAT score of all these students lies between 707.5 to 709.5, and the sample value of 708 is representative of the population parameter. In summary, point estimates such as 708 describes a sample, and confidence intervals tell us what happens in the population and is an estimate of the population parameter. In essence, it provides an estimate of the mean SAT score for all these students (population). The following is the general formula for a centralized confidence interval: $\bar{X} \pm (t)(\text{Standad Error})$. X bar is the mean, and t is the critical value of t for the desired confidence interval, and the standard error is found by finding the standard deviation divided by the square root of the number of scores.

Confidence intervals for coefficient alpha

Confidence intervals for coefficient alpha involve noncentralized distribution. Let us take the example we used before for coefficient alpha. What is the 95% confidence interval around the population coefficient alpha?

The SPSS control lines for this example are the following:

Reliability

```
/VARIABLES=item1 item2 item3 item4 item5 item6
/SCALE('ALL VARIABLES') ALL/MODEL=ALPHA
/STATISTICS=DESCRIPTIVE SCALE CORR ANOVA
/ICC=MODEL(MIXED) TYPE(CONSISTENCY)
CIN=95 TESTVAL=0 .
```

These results for the 95% confidence interval around coefficient alpha were .308 for the lower limit and .897 for the upper limit. Remember from earlier example, the .691 tells us happens with the sample data and is referred to as a sample measure of internal consistency. The 95% confidence interval captures the parameter called the population coefficient alpha, and it means that over repeated samples of confidence intervals, 95% of the intervals will captures the parameter called the population coefficient alpha, and 5% of the intervals will not capture the population coefficient alpha. The 5% chance that values can fall outside of the interval suggests that over repeated samples 2.5% of the intervals will be too low and 2.5% will be too high. In summary, the confidence interval around coefficient alpha deals with the population coefficient alpha that will be represented through several samples or repeated sampling.

It is possible to test a coefficient alpha against a specified value. For example, does a value of .59 differ from the alpha obtained with our coefficient alpha of .691? The SPSS codes for running this analysis are the following:

Reliability

```
/VARIABLES=trial1 trial2 trial3 trial4 trial5 trial6
/SCALE('ALL VARIABLES') ALL/MODEL=ALPHA
/STATISTICS=ANOVA
/ICC=MODEL(MIXED) TYPE(CONSISTENCY)
CIN=95 TESTVAL=.59
```

Results from the F test for the average measures reported an F value of 1.325, p=.236. This indicated that the two values were not statistically significantly different from each other. Testing coefficient alpha against a specific value is an advancement beyond null hypothesis testing. Readers can see Thompson (2003) for a thorough discussion of this advancement in measurement. Finally, confidence intervals can be found for the d effect sizes, and these like coefficient alpha involve noncentralize distributions.

Discussion

This paper addressed three major areas important for hypnosis research. The three issued discussed were measurement, effect sizes, and confidence intervals. Measurement is important for understanding hypnosis research. As previously stated, minorities are seldom included within standardization data for hypnosis. There are over 40 different measures of effect and some are standardized differences like Cohen's *d* or in correlation form like *r*. Finally, effect sizes can be presented as corrected and uncorrected measures.

Thompson (2003) has made a number of recommendations for social sciences research, and this writer thinks the same applies for hypnosis research. He recommended that researchers put confidence intervals around reliabilities like coefficient alpha. As stated within this article, reliability is a function

of test items and reliability measures the consistency of test items. Also, as previously stated, a confidence interval is an interval among an infinitely large set of intervals for a given parameter in which 95% of the intervals would capture the population parameter.

Confidence intervals around reliability indices require a non-centralized distribution – which allows one to a perform power analysis, or the probability of rejecting a false null hypothesis (no treatment effect). The SPSS computer software was used to calculate non-centralized distributions for reliabilities. Unlike centralized distributions, which have a mean of zero, a non-centralized distribution has a mean of some hypothesized value, and non-centralized distributions are skewed (Bird, 2002). As demonstrated within this article, confidence intervals were placed around reliability and validity indices. It should be clear to the reader that in order to construct a confidence interval, one must know the distribution that one is working with such as normal, centralized t-distribution and so on. Confidence intervals allow one to test statistical significance and to find what happens in the population. In contrast, traditional significance testing only allows one to reject or fail to reject the null hypothesis.

I have challenged the use of null hypothesis statistical significance testing within these social sciences (Sapp, 2012; 2006;2017). Readers should be aware that null hypothesis statistical significance only allows one to determine if a relationship is significantly greater than zero, and it does not ensure replication, nor does it control for threats to internal validity.

Internal validity is the judgment applied by a researcher to determine if an independent variable caused a change on a dependent variable, of if hypnosis actually made a difference. Theoretically, random assignment or randomly assigning participants to groups initially controls for all threats to internal validity.

Sapp (2012) recommended that researchers provide effect size measures and reliability indices for their hypnosis data. In addition, he recommended confidence intervals for d effect size measures. Unfortunately, this process is an iterative one that involves non-central distributions and readers who are interested in SPSS programs for calculating such intervals can consult (Bird, 2002; Smithson, 2003). For a nominal fee, Professor Geoff Cummings, at La Trobe University in Australia, has developed software that runs under the Excel program, which can be downloaded from the following website: http://www.latrobe.edu.au/psy/esci. This software calculates confidence intervals for d effect size measures.

Finally, hypnosis researchers need to provide effect size measures for their data, and that they need to calculate reliability indices for their data. In conclusion, hypnosis researchers need to think meta-analytically, and not mindlessly apply statistics and measurement (Fidler, Cumming, Thomason, Pannuzzo, Smith, Fyffe, et al 2005).

Efficacy of hypnosis

Bergin and Garfield (1994) is the definitive source on psychotherapy efficacy. Sapp (1997), citing data from Bergin and Garfield (1994), reported that hypnosis had an average d effect size of 1.82, and he reported a 95% confidence interval around the population d of .8025 for the lower limit and 1.0163 for the upper limit. Four-hundred seventy-five studies were included within this analysis. The r effect size was .68, and this indicated that hypnosis accounted for .4624 of the variance on the outcome measures. The statistical power for this analysis was 1.0, and since statistical power was greater than .90, it was excellent. With cognitive-behavior therapy, Sapp found the point estimate effect size was 1.13, and the 95% confidence interval was .4677 for the lower limit and .6614 for the upper limit. Although the point estimate for cognitive-behavioral therapy had a large effect size of 1.13 (sample effect), the 95% confidence interval found that within the population of cognitive-behavioral therapies, the upper and lower limit fell within the medium effect size range. Byom & Sapp (2013) found that hypnosis had a d effect size of .85 in terms of weight reduction. A 95% confidence interval around the population d was -.34 for the lower limit and 2.04 for the upper limit. The confidence interval indicated that there were not statistically significant differences. The corrected value of d was .78, and a 95% confidence interval around the corrected d was (-.4, 1.96). This confidence interval also found that

there were not statistically significant differences.

In a meta-analysis of hypnosis, Flammer & Bongartz (2003), using 57 studies, found that hypnosis had a weighted or adjusted average effect for d of .56 (medium effect size). For DSM-V diagnosed disorders, they found that hypnosis had a d effect size of .63. They also performed a meta-analysis on 444 hypnosis studies and found a d effect size of 1.07, and the d effect size for randomized studies was .56 and for non-randomized studies a d effect size of 2.29. They found a correlation of .44 between hypnotic susceptibility and treatment outcomes. Like I stated, Flammer and Bongartz (2003) found that effect sizes are needed for hypnosis research, and I believe that confidence intervals are also needed. Clearly, hypnosis is effective, but researchers need to provide data so that reliability and validity indices, effect sizes, and confidence intervals can be calculated.

References

Bergin, A.E., & Garfield, S. C. (eds.) (1994). Handbook of psychotherapy and behavior change *(4th ed.*). New York: Wiley.

Bird, K. D. (2002). Confidence intervals for effect sizes in analysis of variance. Educational and Psychological Measurement, 62, 197-226.

Byom, T. K., & Sapp, M. (2013). Comparison of effect sizes of three group treatments for weight loss. Sleep and Hypnosis, 15(1-2), 1–10.

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

Cohen, J. (1988. Statistical power analysis for behavioral sciences (3rd ed.) New York: Academic Press

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. Professional Psychology: Research and Practice, 40(5), 532.

Fidler, F., Cumming, G. Thomason, N., Pannuzzo, D., Fyffe, P., Edmonds, H.,

Harrington, H. Schmitt, R. (2005). Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology, 73, 136-143.

Flammer, E., & Bongartz, W. (2003). On the efficacy of hypnosis: a meta-analytic study. Contemporary Hypnosis, 20(4), 179-197.

Huberty, C. J. (2002). A history of effect sizes. *E*ducational and Psychological Measurement, *62(2), 227-240*.

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Sapp, M. (2004). Confidence intervals within hypnosis research. Sleep and Hypnosis,6(4), 169-176.

Sapp, M., Obiakor, F.E., Gregas, A., & Scholze, S. (2007). Mahalanobis Distance: A multivariate measure of effect in hypnosis research. Sleep and Hypnosis, 9(2), 67-70.

Sapp, M. (2006). Basic psychological measurement, research designs, and statistics without math. Charles C Thomas Publisher.

Sapp, M. (2017). Primer on effect sizes, simple research designs, and confidence intervals. Charles C. Thomas Publisher.

Sapp, M. (2012). Reliability, Validity, Effects Sizes, and Confidence Intervals in Multicultural Teaching and Learning Research and Scholarship. Multicultural Learning and Teaching, 7 (2).

Sapp, M. (2004). Confidence intervals within hypnosis research. *Sleep and Hypnosis*, 6(4), 169-176.

Sapp, M., Obiakor, F.E., Gregas, A., & Scholze, S. (2007). Mahalanobis Distance: A multivariate measure of effect in hypnosis research. Sleep and Hypnosis, 9 (2), 67-70.

Sapp, M. (2012). Reliability, Validity, Effects Sizes, and Confidence Intervals in Multicultural Teaching and Learning Research and Scholarship. Multicultural Learning and Teaching, 7(2).

Sapp, M. (1997). *Counseling and psychotherapy: Theories, associated research, and issues.* Lanham, MD: University Press of America.

Sapp, M. (2010). *Psychodynamic, affective, and behavioral theories to psychotherapy*.Charles C Thomas Publisher.

Steiger, J. H., & Fouladi, R. T. (1997). *Noncentrality interval estimation and the evaluation of statistical models.* In L. Harlow, S. Mulaik, & J. Steiger

(Eds.), What if there were no significance tests? (pp. 221-256). Mahwah, NJ: Erlbaum.

Smithson, M. (Ed.). (2003). *Confidence intervals* (No. 140). Sage.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.).Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, 31(3), 25-32.

Thompson, B. (Ed.). (2003). Score reliability: Contemporary thinking on reliability issues. SAGE Publications, Incorporated.

Smithson, M. (Ed.). (2003). Confidence intervals (No. 140). Sage.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.).Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, *31*(3), 25-32.

Thompson, B. (Ed.). (2003). Score reliability: Contemporary thinking on reliability issues. SAGE Publications, Incorporated.